# Augmenting Text Generation Approaches in Question Generating Transformer Models

**Intae Kim  Steve Hewitt**
University of California, Berkeley

## Abstract

This paper outlines the application of augmentative text generation and decoding strategies for transformer-based neural networks to the question generation task for the SQuAD 1.1 dataset. The primary augmentations employed are the usage of Gaussian Process Priors in between the encoder and decoder of the T5-base model, and also nucleus sampling in the decoder, which have been shown to improve evaluative metrics and diversity in text generation. While the augmented models discussed were unable to improve the overall evaluation metrics of SacreBLEU, ROUGE, and METEOR, noteworthy spikes in performance in the augmented models suggest further research should be committed to this research avenue.

## 1 Introduction

This paper outlines the application of augmentative text generation and decoding strategies for transformer-based neural networks to the question generation task for the SQuAD 1.1 dataset. The primary augmentations employed are the usage of Gaussian Process Priors in between the encoder and decoder of the T5 model, and nucleus sampling in the decoder, which have independently been shown to improve evaluative metrics and diversity in text generation. While the augmented models discussed were unable to improve the overall evaluation metrics of SacreBLEU, ROUGE, and METEOR, noteworthy spikes in performance in the augmented models suggest further research should be committed to this research avenue.

## 2 Dataset

The dataset used for training purposes was a subset of the Stanford Question Answering Dataset (Rajpurkar et al. 2016), which contains 107,785 question-answer pairs on 536 articles generated through crowdsourcing question-answer pairs based on Wikipedia articles. SQuAD was designed for question answering tasks but the part of the dataset that is publicly available has been frequently repurposed for question generation (Xiao et al. 2021) (Lopez et. al 2021) by a variety of researchers. SQuAD offers diversity in content and question formats, which promotes greater relevance and fluency in question generation on unforeseen tasks. It avoids specialized question types like true or false and multiple choice, instead focusing on short open-response question and answer pairs. Because the test set for SQuAD is not publicly available, a test set must be created from a portion of the training or validation set. We elected to follow the technique that was more successful in the ERNIE-GEN paper and randomly sample 50% of the validation set for use as our test set.

## 3 Background

**Question Generation:** Machine question generation is a less well-studied topic than question answering, but there were several prominent papers on the topic that informed our work. ERNIE-GEN (proposed by Xiao et al. 2020) was the top-performing QG model that we could find. Their model had several interesting features including span-by-span generation to improve on grammatical consistency over word-by-word generation, noise-aware generation from pre-training on intentionally corrupted data, and a technique called a multi-granularity target

fragments that the authors assert forces the encoder to rely more on the encoder layers' output than on the previously generated words when completing a sequence. Although we intended to work with the ERNIE-GEN framework, we abandoned this effort due to the lack of English-language reference materials. We also looked at the results from the Lopez et al. (2021) paper on question generation, especially for comparison of scoring metrics and dataset preprocessing.

**Gaussian Process Priors:** Gaussian Process Priors have recently been demonstrated to improve both the primary metric score and diversity of answers for tasks like paraphrase generation. This method transforms the deterministic states from a transformer's encoder's hidden layers into random hidden states using a stochastic function. This introduces variability while preserving the encoded context. An attention calculation is applied to these random hidden states to produce context vectors that feed into the model's decoder at each time step. These steps have the effect of allowing for generating a word sequence conditioned on previously sampled words, rejecting the approach utilized by other traditional methods which consider all words to be independent of each other (Du et al., 2022). The novelty of our approach comes from applying a GPP model to the question generation task. We hypothesized that the improvements seen in other sequence-to-sequence tasks achieved by implementing a GPP-enabled T5 model could possibly extend to the question generation task, which is substantially similar.

**Beam Search:** Beam search is a method of surveying the output space by keeping the top partial sequences at each time step, as ranked by their conditional probability, and ending with the complete sequence(s) with the highest final conditional probability amongst the remaining candidate beams. Beam search is a computationally efficient method, as it searches a subset of the joint probability distribution of the decoder for the optimal solution, and this efficiency has helped drive its popularity. However, it is a greedy algorithm as only a preset number of the "best" solutions are kept at each step. Though it is a fundamental improvement over greedy search, which is equivalent to beam search with a beam size of 1, there is no guarantee that this method will produce the optimal solution.

While beam search is an effective method for the time and computation complexity required, there are augmentative methods that have grown in recent popularity that may be more effective at surveying the output probability space, generating more diverse text, and producing more human-like text. In this paper, we will implement a number of recently developed methods such as Gaussian Process Priors and nucleus sampling in an attempt to improve on previously top results obtained by T5 for question answering on SQuAD 1.1.

**Nucleus Sampling:** Also known as top-p sampling, nucleus sampling selects the minimum number of words whose combined probability exceeds that of a predetermined probability threshold. Through this method, we can focus on the most relevant set of the probability distribution, referred to as the nucleus. Furthermore, we are able to dynamically adapt the size of the sampled set within the nucleus at each time step whereas the aforementioned method of beam search had a fixed output sequence size. With this added flexibility in approach, recent papers have found that this decoding strategy has resulted in improved text generation diversity and fluency (Shaham et al., 2021).

## 4 Models

**Baseline T5:** T5 is a text-to-text transformer model which is capable of being applied to any NLP task. As the model is at its core based in transfer learning, supported by the architecture which both accepts input and output as a text string, T5 utilizes an enormous amount of unlabeled data to generalize for text prediction tasks. We're then able to leverage the pretrained model to fine-tune the model on smaller labeled datasets like the SQuAD 1.1 question-answer pair dataset to optimize performance for our particular task of question generation.

**Experimental T5 with GPP:** For our GPP implementation of T5, we look to borrow the work of Du et al. (2022), which utilizes Gaussian process priors to generate random context variables between the encoder and decoder to encourage text generation diversity. In their paper, they find that T5 adapted with a GPP implementation improves self-BLEU, which is a measure of token-level repetition. This is consistent with the expectation that using GPP for text generation generates more

diverse results, though they found that overall quality metrics were sometimes degraded in consequence.

## 5 Methodology

**Evaluation Metrics:** To evaluate our models, we used a combination of SacreBLEU, ROUGE, and METEOR. SacreBLEU is an implementation of traditional BLEU that differs slightly in tokenization rules, but fulfills a similar purpose in measuring the similarity between a hypothesis and reference text. It additionally differs from ROUGE in that while both measures capture the precision and recall for a particular text pair, BLEU imposes a brevity penalty that is not present in the ROUGE metric. METEOR focuses on the presence, alignment, and ordering of unigrams in both the hypothesis and reference text. These metrics all score the similarity of the model output to the target text, and they were selected because of their prevalence in related academic works about sequence-to-sequence modeling. Each of these metrics was designed to grade machine translation tasks, but have been adopted to evaluate most text-generation tasks by the machine learning community. Our research into question generation showed these same metrics were used by the ERNIE-GEN team (Xiao et. Al 2019) and the cited GPT-2 implementation (Lopez et. al 2021). BLEU and METEOR were the metrics used to evaluate the GPP models (Du et. al 2022), although they did not test their model on a question generation task.

**Implementation:** For our experiment we chose to implement two main model architectures: our baseline T5-base model, and a T5-base model modified to make use of Gaussian Process Priors (hereafter referred to as GPP). To ensure a fair comparison between models we needed to align the hyperparameters and training conditions as closely as possible. One of the major differences between baseline T5 and T5 with GPP is the GPP version takes considerably more time to train, and this difference grows exponentially with the maximum length of the encoder. The GPP model is slower because it has to calculate the full covariance matrix of context variables of the GP prior during training. During inference the difference in calculation time between baseline and T5 with GPP is trivial. We considered encoder lengths of 64, 128, 256, and 512 tokens and determined that the 256-length version would maximize the amount of context being fed into the model without making training prohibitively time-consuming.

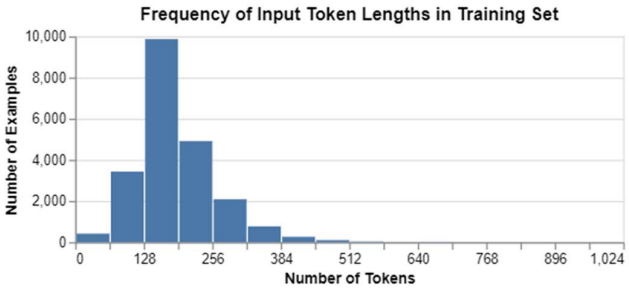



Figure 1: Histogram of Input Token Lengths

Due to the training time required and the limited hardware available, we elected to reduce the training size by 75% to 21,900 examples, and implement early stopping points for some of the larger models. However, as the greatest improvements in loss occurred earlier on in training, it's unlikely that this restriction caused significant degradation in model training and performance, which is also supported by our results discussed in a later section. Without reducing the training set, our GPP model would have taken an entire month to train or longer and this would have prevented testing variations of the design. We based our hyperparameters on a combination of the default values from the GPP codebase and the configuration that the ERNIE-GEN team used for question generation on the same dataset.

To test the viability of a GPP-based approach, we originally trained a smaller GPP model to see if the results were comparable to our baseline. This model achieved a BLEU score of 15.19, which was below our expectations, so it was sent for additional training. Despite the loss function showing improvement against the validation set, the final version of this smaller model performed even worse than the previous iteration, with a BLEU score of 4.69. This demonstrated that a model could degrade in BLEU score while still improving in loss. The outputs of the model showed some discouraging behavior such as repeating phrases and allowing the answer to contaminate the question. We hypothesized that the model was reinforcing bad habits when overtrained, perhaps fitting to the noise that is inherent in a non-deterministic output. This problem led us to perform hyperparameter tuning

by grid search to identify the optimal values for the GPP-exclusive kernel V and kernel R, and to capture additional checkpoints in our final model in case similar degradation occurred.

| Hyperparameter | Value |
|---|---|
| Encoder Max Length | 256 |
| Decoder Max Length | 48 |
| Maximum Return Sequence Length | 96 |
| GPP Kernel V | 100 |
| GPP Kernel R | 0.001 |
| Beam Size (during training) | 5 |
| Batch Size | 8 |
| Number of Epochs | 5 |
| Learning Rate | 0.0001 |

Table 1: Hyperparameters used during model training

Additional grid search optimization was performed around the hyperparameters needed for beam search and the combination of nucleus and top-k sampling. Our experiments showed that a beam size of 5 is optimal for this task for both our T5 baseline and GPP model against the validation set. Nucleus search and top-K gave their best results against the validation set with a top-P of 0.5 and top-K of 100 used in concert. It was more difficult to gauge the value of nucleus search for the GPP version, because the outcome of the model is not deterministic. Certain nucleus sampling hyperparameters showed BLEU score improvements of 15% or more over beam search during our grid search, but the final result showed nucleus sampling to be a hindrance to model performance overall.

## 6 Experiment Results

**Overall Performance:** Following training and grid search optimization of beam search and nucleus search hyperparameters, predictions against the test were generated and scored for both models and both decoding strategies. The Baseline T5 model with Beam Search performed better than all other versions as measured by BLEU (21.13), ROUGE-L (46.28), and METEOR (47.45) scores. For both our Baseline model and our GPP model we saw a degradation in performance when using nucleus search instead of beam search.

We also graded every individual prediction and compared BLEU scores across models to determine if any of the models had made a perfect prediction (exactly matching the target), which model had the best prediction on each item, and which model had the worst prediction on each item. Best and worst distinctions were only made in instances where multiple models did not independently arrive at the same best or worst output. One type of error that is particularly harmful to the task of question generation, but not captured in any of our primary metrics, is answer contamination. Generating a question that includes the answer can earn a high score on the other metrics but is essentially useless for practical purposes. The GPP model showed 13.26% contamination using beam search, and 8.38% contamination using nucleus search, compared to less than 1% contamination from the Baseline model using either strategy. Even if the GPP model had scored higher than the Baseline model in every other metric, this level of answer contamination would suggest that the Baseline model is still superior. Other research (Lopez et. Al 2021) suggest that answer-aware question generation does not necessarily benefit from knowledge of the answers unless the model is designed with the

| Model Version | BLEU | ROUGE-L | METEOR | Perfect Predictions | Best Predictions | Worst Predictions | Answer Contamination |
|---|---|---|---|---|---|---|---|
| Baseline T5 Beam Search (beam size = 5) | **21.13** | **46.28** | **47.45** | **3.22%** | **19.53%** | **11.28%** | **0.76%** |
| Baseline T5 Nucleus Search (p = 0.5, k = 100) | 18.26 | 44.26 | 43.90 | 2.61% | 16.50% | 14.46% | 0.79% |
| GPP T5 Beam Search (beam size = 5) | 16.57 | 40.19 | 41.43 | 2.01% | 15.61% | 22.25% | 13.26% |
| GPP T5 Nucleus Search (p = 0.5, k = 100) | 13.64 | 37.32 | 37.42 | 1.19% | 14.87% | 28.33% | 8.38% |

Table 2: Results from finalized models predicting against the test set. Perfect, Best, and Worst predictions are determined by BLEU score


Figure 2: BLEU score differential by tokenized context length between each prediction set

answer-awareness in mind or learns how to attend to the answer during training.

**Error Patterns**: Even with the poor performance in the three key metrics and all of the answer contamination, it is interesting to see that each of the four prediction sets generate the best prediction of the group for 14.87% to 19.53% of the test items. We looked to identify patterns in the score differential between the models and analyzed examples where each version had the best prediction without being a perfect match to the target. As the length of the tokenized context increases, the nucleus search version of both models performs worse when compared to the beam search version. Other works that we examined on the topic of question generation

| Context | Target | Answer | Good Model | Good Prediction | Bad Model | Bad Prediction |
|---|---|---|---|---|---|---|
| ...Also **on loan to the museum**, from Her Majesty the **Queen Elizabeth II**, **are the Raphael Cartoons**... | Who has loaned the Raphael Cartoons to the museum? | Queen Elizabeth II | Baseline (Beam) | **Who loaned the Raphael Cartoons to the museum?** | GPP (Beam) | Who was the **Queen Elizabeth II**? |
| ...Another important monument, **the statue of Little Insurgent** located at the ramparts of the Old Town, **commemorates** the children **who served as messengers and frontline troops in the Warsaw Uprising**... | Who does the statue of Little Insurgent commemorate? | children | GPP (Beam) | What did **the statue of Little Insurgent commemorate ?** . | Baseline (Beam) | Who **served as messengers and frontline troops in the Warsaw Uprising?** |
| ...**Harvard Museum of Natural History** includes the Harvard Mineralogical Museum, Harvard University Herbaria featuring the Blaschka Glass Flowers exhibit, and the Museum of Comparative Zoology... the Peabody Museum of Archaeology and Ethnology, **specializing in the cultural history and civilizations of the Western Hemisphere**... | What museum specializes in cultural history and civilizations of the Western Hemisphere ? | Peabody Museum of Archaeology and Ethnology | Baseline (NS) | **What museum specializes in** the **cultural history and civilizations of the Western Hemisphere?** | GPP (NS) | **What museum** does **the Harvard Museum of Natural History** have? |
| On May 1, 1953, **ABC's New York City flagship stations** – WJZ, WJZ-FM and WJZ-TV – **changed their** respective **callsigns** to **WABC, WABC-FM and WABC-TV**... | When did ABC's New York flagship stations change their call signs? | May 1, 1953 | GPP (NS) | **When did ABC's New York flagship stations change their callsigns to**? | Baseline (NS) | **When did ABC change their callsigns to WABC, WABC-FM and WABC-TV?** |

Table 3: Sample sentences displaying significant discrepancies in score by model

spent little to no time discussing the types and causes of errors their models made.

The sample sentences table shows examples where one model scored much higher than a counterpart and highlights the parts of the context that informed each model's output. The first example shows answer contamination from the GPP model. The second example shows how the Baseline model ended up with a poor score even though it generated a question that fit with both the context and the answer. The third example shows the GPP model producing a nonsensical question and referencing the incorrect museum instead of the one designated to be the answer. The fourth example shows the GPP model getting a high score for a question that has grammatical issues, while the Baseline received poor scores for a question that seems to be entirely valid given the context and answers. Collectively these samples illustrate how the Baseline model is often given poor scores despite returning what is a valid question, how the GPP model sometimes gets better scores when returning grammatically incorrect questions, and the tendency of answer contamination in the GPP output. These patterns were repeated in many additional examples that we reviewed manually, and in novel examples that we created outside of the SQuAD dataset.

## 7 Conclusion

Despite our best efforts, we were unable to improve the performance of our Baseline T5 model with beam search by implementing gaussian process priors, nucleus search with top-k, or a combination of the two. The task of question generation may not be a suitable application for the added variety of text output that a GPP model provides. The brevity of the target output sequence did not allow much room for improvisation. The tasks that the team behind the original GPP paper (Du et Al. 2022) targeted such as paraphrase generation and text style transfer seem better suited to a machine learning model that leans towards exploring new corners of the probability space rather than exploiting the well-worn pathways that deterministic models settle into.

Given more time and resources, there are a number of research avenues left unexplored that we'd like to investigate in the future including developing a hybridized loss function to discourage answer contamination, using the unabridged SQuAD 1.1 training set, including multiple QG datasets in our training material, and implementing a GPP model with the maximum 512-length encoder.

## References


Du, W., Zhao, J., Wang, L. and Ji, Y., 2022. *Diverse Text Generation via Variational Encoder-Decoder Models with Gaussian Process Priors*. arXiv:2204.01227.

Holtzman, A., Buys, J., Du, L., Forbes, M. and Choi, Y., 2019. *The curious case of neural text degeneration.* arXiv:1904.09751.

Lopez, L.E., Cruz, D.K., Cruz, J.C.B. and Cheng, C., 2021, November. Simplifying paragraph-level question generation via transformer language models. In *Pacific Rim International Conference on Artificial Intelligence* pages 323-334. Springer, Cham.

Rajpurkar, P., Zhang, J., Lopyrev, K. and Liang, P., 2016. *Squad: 100,000+ questions for machine comprehension of text*. arXiv:1606.05250.

Shaham, U. and Levy, O., 2021. *What Do You Get When You Cross Beam Search with Nucleus Sampling?*. arXiv:2107.09729.

Xiao, D., Zhang, H., Li, Y., Sun, Y., Tian, H., Wu, H. and Wang, H., 2020. *ERNIE-GEN: an enhanced multi-flow pre-training and fine-tuning framework for natural language generation*. arXiv:2001.11314.


## Other References


Du, Wanyu et al. 2022. "GP-VAE" [online at: https://github.com/wyu-du/GP-VAE, accessed July 2022].

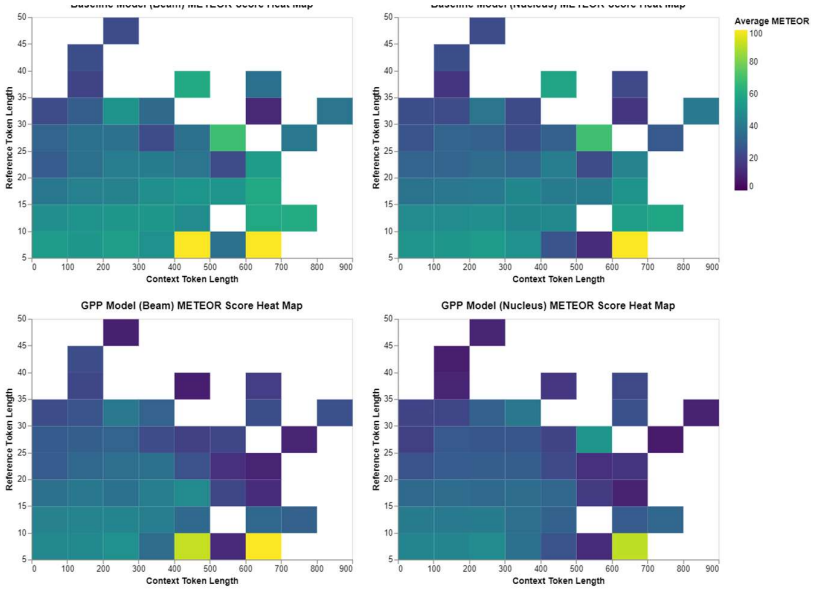
Huggingface.co. 2022. "How to generate text: using different decoding methods for language generation with Transformers." [online at: https://huggingface.co/blog/how-to-generate, accessed July 2022].

Paperswithcode.com. 2022. "Papers with Code - SQuAD1.1 Benchmark (Question Generation)." [online at: https://paperswithcode.com/sota/question-generation-on-squad11, accessed July 2022].

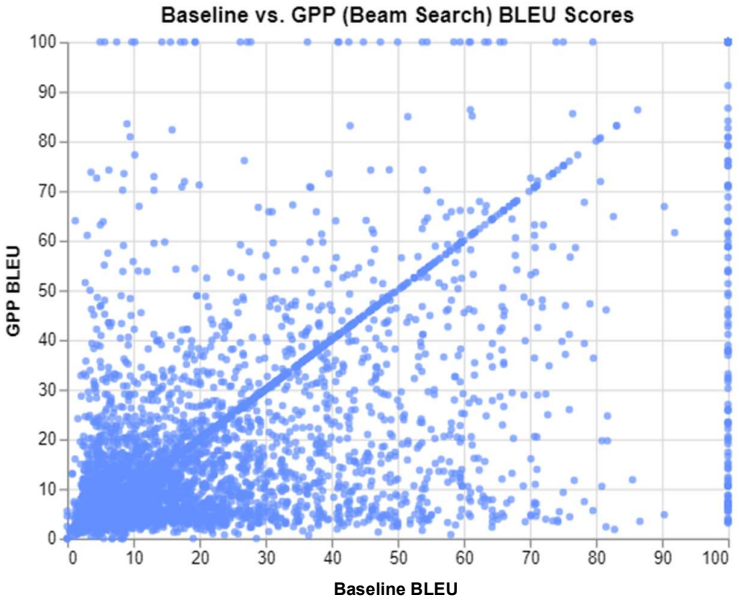
PaddlePaddle. 2022. "ERNIE: Official implementations for various pre-training models of ERNIE-family, covering topics of Language Understanding & Generation, Multimodal Understanding & Generation, and beyond." [online at: https://github.com/PaddlePaddle/ERNIE, accessed July 2022].

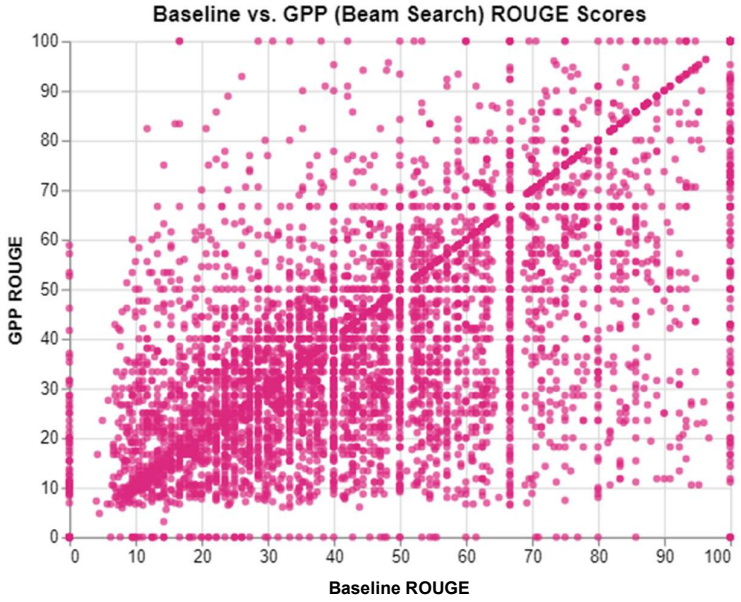
## Appendix

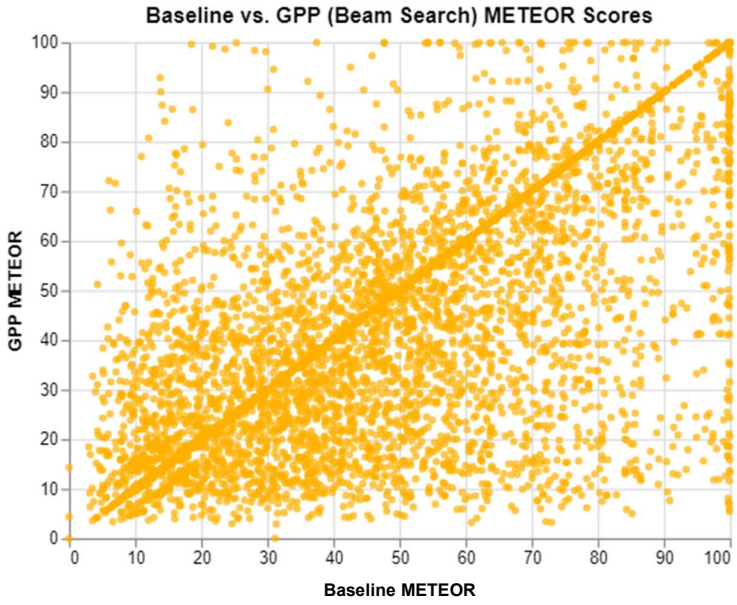
Appendix 1. 1: Prediction METEOR Score Heatmap




Appendix 1. 2: BLEU, ROUGE-L, and METEOR score comparison between Baseline and GPP Beam Search predictions on the test set.




Baseline vs. GPP (Beam Search) ROUGE Scores
GPP ROUGE
Baseline ROUGE
0
10
20
30
40
50
60
70
80
90
100


Baseline vs. GPP (Beam Search) METEOR Scores
GPP METEOR
Baseline METEOR
0
10
20
30
40
50
60
70
80
90
100