Believable Story Generation using Transformers

Rahul K Johny
MSc in Computing (Artificial Intelligence)
Dublin City University
Student No. 22262108
rahul.kuruppasseryjohny2@mail.dcu.ie

Rohit Shinde
MSc in Computing (Data Analytics)
Dublin City University
Student No. 22260213
rohit.shinde2@mail.dcu.ie

Supervisor: Dr. Jennifer Foster
Faculty of Engineering and Computing
Dublin City University
Assistant Professor
jennifer.foster@dcu.ie

***Abstract*—This study looks into the narrative generation capabilities of OpenAI's GPT-2, GPT-3, and GPT-4 language models. We analyse created narratives using BERTScore and Self-BLEU, augmented by human evaluations, through fine-tuning and zero-shot experiments. The results show that fine-tuned models produce higher-quality narratives, with zero-shot GPT-4 having equivalent capabilities. While automatic measurements correlate with human scores, the correlation is larger for lower-rated stories, showing that automatic analytics have limits in capturing nuanced narrative quality. These findings provide important insights into these models' content generation capabilities, directing future research.**

***Index Terms*—Natural Language Processing (NLP), Fine-tuning, Zero-shot Learning, GPT-2, GPT-3, GPT-4, Story Generation, Transformers.**

I. Introduction

Artificial intelligence's proficiency in language understanding and generation has transformed how we interact with technology. At the heart of this revolution lies the field of natural language processing (NLP), and more specifically, text generation. This area has seen remarkable growth and innovation, pushing the boundaries of what machines can achieve in emulating human-like communication. The primary objective of this research is to delve into this fascinating domain, investigating the capabilities of state-of-the-art language models, namely GPT-2 and GPT-3, in generating compelling, believable narratives. Creating an engaging story is more than assembling well-structured, grammatically correct sentences; it calls for the elusive trait of creativity, commonly considered an exclusive hallmark of human cognition. Incorporating this human-like imagination into computational models is a complex, yet exciting challenge, one that could significantly expand the potential applications of text generation. While the recent progress in NLP provides promising starting points, a comprehensive exploration of this capability is still lacking.

In this study, we have examined the effectiveness of GPT-2, GPT-3, and GPT-4 in generating narratives that exhibit coherence, relevance, creativity, and engagement. Our results suggest that fine-tuning can enhance the narrative generation abilities of these models, while the newest, larger models such as GPT-4 demonstrate competitive performance even in a zero-shot scenario. The integration of human-like storytelling into AI models has significant implications for the field of content creation, demonstrating potential for both text-based narratives and, as future work may explore, generating prompts for other AI-driven media.

II. Background Work

Stories require uniqueness, a high-level storyline, and thematic consistency throughout the text, which involves modelling extremely long-term connections. [1] is a highly influential paper in the field of NLP, with an impressive citation count exceeding 50,000 which introduces the Transformer neural network architecture. Attention plays a crucial role in human cognitive activities, helping us filter and prioritise relevant information. Various fields, including philosophy, psychology, neuroscience, and computer science, have explored the concept of attention. Deep neural networks have recently focused on studying attention [5]. Before [1], machine translation architectures predominantly relied on encoder-decoder structures, using recurrent neural networks (RNNs) to capture language sequences. However, RNNs had limitations in capturing long-range dependencies and faced challenges in parallelizing computations due to data race conditions. In 2014, Bahdanau et al. [6] introduced the concept of "attention," which improved intermediary representations by incorporating contextually important words during the encoding process [6]. This attention mechanism was further utilised by Vaswani et al. [1], where the Transformer model architecture was introduced. Transformers employ self-attention, allowing words to associate with each other at a high level and enabling accurate referencing of pronouns such as "it." Vaswani et al. [1] were able to significantly improve model performance and calculation speed by removing recurrence from a well-established encoder-decoder design.

Zhang et al. [2] introduced BERTScore as an automatic evaluation metric specifically designed for text generation tasks. In their study, they utilized BERTScore to evaluate the performance of 363 different systems, which included machine translation and image captioning models. The cosine similarities between the token embeddings of two sentences are added up by BERTScore to determine how similar they are. They used an n-gram matching approach and several metrics such as BLEU, METEOR, RUSE, NIST, etc in their work. With reference sentence $x = (x_1, \ldots, x_k)$ for ex. "the weather is cold today" and a candidate sentence $\hat{x} = (\hat{x}_1, \ldots, \hat{x}_k)$ for ex. "it is freezing today", authors used contextual embeddings to represent the tokens, and compute

matching using cosine similarity, optionally weighted with inverse document frequency scores [2] [15]. Writers compared the results of machine translation and image captioning systems in which BERTScore outperforms other metrics. Also it resolves some of the existing problems in other metrics. The authors conducted numerous experiments with various setups, and they found that BERTScore outperformed existing metrics in terms of model selection and had a higher correlation with human judgements.

Akoury et al. [3] have used segment embeddings to fine-tune a pre-trained language model (GPT-2) on the [1]STORIUM dataset to differentiate each sort of context. One of the key issues of GPT-2 is: any entry in a story may be dependent on any narrative element that comes before it (e.g., previous entries, scenes, challenges) [4] [9]. They attempted to address this issue by offering a platform for testing models in a machine-in-the-loop situation by allowing genuine STORIUM authors to engage with the created stories. However, due to the effort that STORIUM authors have to put into reviewing the stories, as well as the STORIUM community's small size, evaluations can take a very long time. As a result, they determined that the platform is not yet ready for immediate review of generated content. Furthermore, because the assessment platform is only available on STORIUM, it cannot be used to test models trained on other story-generating datasets, as the website's users are primarily interested in writing narratives in the STORIUM format.

Guan et al.[4] also uses transformers, specifically GPT-2 for generating stories. They have taken a different approach to keep the content generated within the same context. They improved GPT-2 with such knowledge by post-training the model on knowledge examples built from these knowledge bases, which can provide additional critical information for story production. Experiments show that training with millions of instances increases generated stories' coherence and logicality [10], "Coherence" refers to stories that are understood and flow naturally, whereas "logicality" refers to stories that are more sensible and follow a logical order of events. In the meantime, they used multi-task learning to solve the difficulty of dealing with cause and temporal connections. They combined the generation objective with an auxiliary multi-label classification objective that requires identifying authentic stories from false stories created by randomly mixing sentences, substituting a sentence with a negatively sampled text, or repeating a sentence in an original story. Through their tests, they demonstrated that the proposed approach uses both implicit knowledge from deep pre-trained language models and explicit knowledge from post-training on external commonsense knowledge bases, resulting in improved performance for a commonsense narrative generation. To assess the flow and logic of the generated stories, the authors use manual evaluation. They randomly sampled 200 stories from the test set for manual evaluation and acquired 1,800 stories from the nine models. Three annotators were engaged for each pair of stories (one by their model and the other by a strong baseline model) to give a preference (win, lose, or tie) in terms of two metrics. For annotation, they used the crowdsourcing service Amazon Mechanical Turk (AMT), and they used a majority vote to make final selections among the three annotators. The results show that their model outperforms other baselines significantly. Post-training on knowledge bases, in particular, leads to major improvements in grammar and logic by providing more knowledge for expanding story plots. Furthermore, multi-task learning improves reasoning performance while having little effect on the fluency of generated stories.

Zhu et al. [11] introduced Texygen, a versatile platform designed for implementing various text generation models and evaluating the diversity, quality, and consistency of the generated texts. They conducted experiments using baseline models such as Vanilla MLE, SeqGAN, MaliGAN, RankGAN, GSGAN, TextGAN, LeakGAN, among others, and evaluated the results using metrics including BLEU, EmbSim, NLL-oracle, NLL-test, and Self-BLEU. One notable metric used in their experiments is Self-BLEU, which evaluates the diversity of the generated text by measuring the similarity between sentences. By treating one sentence as a hypothesis and the others as references, the authors calculated the BLEU score for each generated sentence and defined the average BLEU score as the Self-BLEU of the document [11]. The experiments conducted with the baseline models and evaluation metrics showed that Self-BLEU yielded promising results, with the MLE baseline model achieving the lowest score. The use of Self-BLEU as an evaluation metric is advantageous in text generation tasks because it does not rely on a gold standard or reference text. Instead, it leverages the generated sentences themselves to assess their diversity and dissimilarity. This allows for a more flexible and self-contained evaluation process, which is particularly useful when a gold standard text is unavailable or difficult to define.

III. Methodology

A. How GPT Works

Transformers have proven to be highly effective in text generation tasks. The self-attention mechanism employed in Transformers allows them to capture contextual relationships between words, enabling the generation of coherent and contextually relevant text. Unlike traditional recurrent neural networks (RNNs), Transformers can process information in parallel, making them well-suited for capturing long-range dependencies in text.

They are based on an encoder-decoder structure, where the encoder maps an input sequence of symbols to a sequence of continuous representations, and the decoder generates an output sequence of symbols one element at a time, using the continuous representations produced by the encoder.

The encoder and decoder are composed of a stack of identical layers. Each layer in the encoder has two sub-layers: a multi-head self-attention mechanism and a position-wise fully connected feed-forward network. The decoder has an additional third sub-layer that performs multi-head attention

[1]https://storium.cs.umass.edu/

over the output of the encoder stack. Residual connections are employed around each of the sub-layers, followed by layer normalization. This helps in training deeper models by mitigating the problem of vanishing gradients.

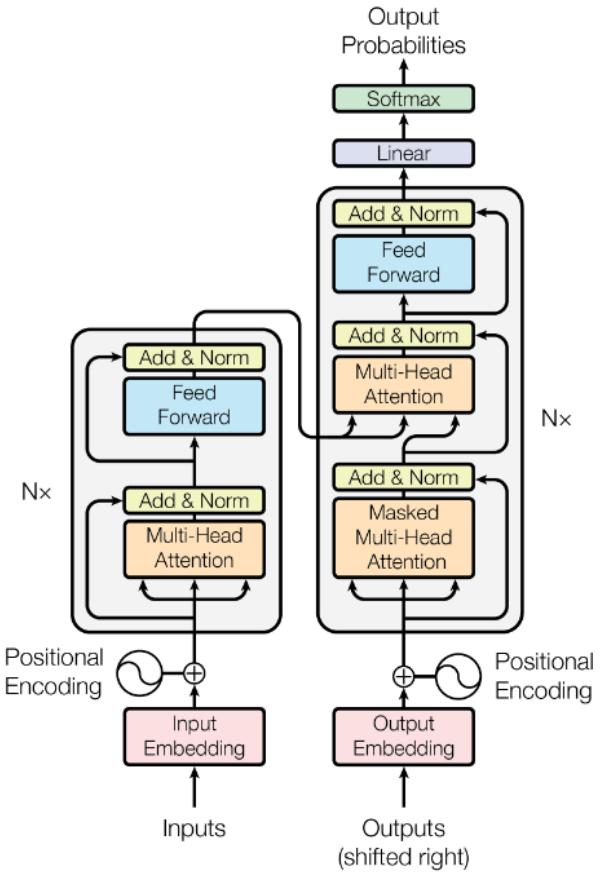

Fig. 1. The Transformer architecture, developed by Vaswani et al. [1]

GPT-2 (Generative Pre-trained Transformer 2) is a prominent example of a Transformer-based language model used for text generation. It has been pre-trained on a vast corpus of text data and can generate high-quality text by predicting the next word given a sequence of input words [12]. GPT-2 has demonstrated impressive abilities in various text generation tasks, such as story writing, poetry generation, and dialogue generation. Its sophisticated language modelling capabilities allow it to generate human-like text that exhibits coherence and relevance.

Building upon the success of GPT-2, GPT-3 (Generative Pre-trained Transformer 3) represents a further advancement in text generation. With an even larger number of parameters, GPT-3 exhibits unprecedented capabilities in generating diverse and contextually accurate text. GPT-3 has demonstrated remarkable proficiency in various text generation applications, including content creation, code generation, and chatbot interactions. Its ability to understand prompts and generate responses in a human-like manner has garnered significant attention in the field [13].

Self-Attention is a fundamental component of Transformers, GPT-2, and GPT-3 that enables them to capture relationships between different words in a text [6]. It allows the models to weigh the importance of words based on their contextual relevance within the given input. To understand self-attention, let's consider the sentence "The cat sat on the mat". In a self-attention mechanism, each word in the sentence is represented as a vector. These vectors are then used to calculate attention scores between words [1]. For example, when processing the word "cat," the model calculates attention scores with respect to all other words in the sentence: "the", "sat", "on" and "mat" but also evaluates the importance of the word "cat" itself. The attention scores reflect the relevance or importance of each word in understanding the context of "cat." Higher attention scores indicate that a particular word is more relevant to the current word being processed. In this case, "cat" might have higher attention scores with "the" and "mat" since they provide important contextual information. On the other hand, the attention scores between "cat" and "on" might be lower since they are less relevant to each other. By calculating attention scores for each word pair in the sentence, the model captures the relationships and dependencies between words. This allows the model to understand the context and generate appropriate responses in text generation tasks.

In summary, self-attention mechanisms enable Transformers, GPT-2, and GPT-3 to capture the relationships and dependencies between words in a text. This helps the models understand context and generate coherent and contextually relevant text.

B. Hyperparameter Fine-Tuning

Hyperparameter tuning is a crucial step in optimising the performance of GPT-2 and GPT-3 models for text generation tasks. We focused on tuning three key hyperparameters: temperature, max_tokens, and top_p. Temperature controls the randomness of the generated output, with higher values resulting in more diverse but potentially less coherent text. Max_tokens limits the length of the generated text, ensuring it remains within a specified token limit. Top_p, also known as nucleus sampling, controls the diversity of the generated output by restricting the probability distribution to a subset of the most likely tokens [14]. To identify the best settings for these hyperparameters, we conducted a grid search, exploring various combinations of values and evaluated the generated stories using our predefined evaluation metrics. This iterative process allowed us to fine-tune the hyperparameters and select the configuration that produced the highest quality and most engaging stories.

C. Evaluation

1) ***BERTScore****:* BERTScore computes a similarity score for each token in the candidate sentence with each token in the reference sentence [2]. BERTScore focuses on computing semantic similarity between tokens of reference and hypothesis. The idea is to understand the meaning of what is generated and what was supposed to be generated and then

comparing the generation against the ground truth. There are two sentences, reference sentence (x) and candidate sentence ($\hat{x}$) which is passed through a pre-trained BERT model that generates contextual embedding for each of the words at the output end and when we have final embedding for each of these words we do n square computation by calculating similarity for each of the words from reference to each of the words in the candidate set. Then similarity metrics will have all the relations of words with each other. We sum the maximum similarity for each of the words from the reference to the candidate and then normalise it, which we call recall. The mentioned process involved normalizing the candidate's data with respect to the reference context, specifically by dividing it by the total number of words in the candidate. This metric is referred to as precision. We can then calculate the f1-score, which is the harmonic mean of precision and recall.

*2) **Self-BLEU**:* Self-BLEU is a variant of the BLEU score that is used to evaluate the diversity of generated sentences in a text generation model [11]. It is particularly useful in scenarios where a model generates multiple sentences, and we want to ensure that these sentences are not just high-quality but also diverse. The basic idea behind Self-BLEU is similar to BLEU. It assigns a numerical score to a set of generated sentences that indicates how similar they are to each other. Instead of comparing the generated sentences to a set of reference sentences, Self-BLEU compares each generated sentence to all other generated sentences. Like BLEU, Self-BLEU also uses n-grams to compare sentences. An n-gram is a contiguous sequence of n tokens from a given sample of text or speech. Self-BLEU calculates the score by taking the average number of common words in each generated sentence and the total number of words in the sentence. For example, if we have two generated sentences: "I have forty six years" and "I am forty six years old", the Self-BLEU score would be calculated based on the common n-grams in these two sentences. The precision would be the number of common n-grams divided by the total number of n-grams in the sentence. However, like BLEU, Self-BLEU also faces the issue of repetitive words leading to inflated precision scores. To address this, Self-BLEU also uses a modified precision that clips the number of times a word is counted, based on the maximum number of times it appears in the other generated sentences. In addition, Self-BLEU also considers the order of words by computing the precision for several different n-grams. The final Self-BLEU score is then the geometric mean of these precision scores, which ensures that the generated sentences are not just diverse at the word level, but also in terms of their overall structure and content [11].

*3) **Human Evaluation**:* In our research, we used an effective human evaluation approach to assess the quality of stories generated by the GPT-2, GPT-3, and GPT-4 models. Recognizing the limitations of automatic evaluation metrics, we have supplemented our use of BERTScore and Self-BLEU with a comprehensive human evaluation process. To begin with, we selected the top-performing stories from each of our fine-tuned models based on their BERTScore and Self-BLEU evaluations. We assume that these stories represent the high quality outputs from each model, as determined by our automatic evaluation metrics. We then developed a [2]survey to facilitate the human evaluation process. We showed each story to the reviewers without telling them which model created it, to avoid any bias. We asked the reviewers to score each story based on how interesting, relevant, creative, and engaging it was, using a scale from 1 to 10. This way, we could get detailed feedback on each story. We also gave the reviewers a chance to write down their thoughts about each story. This let them share any extra insights that the scoring system might have missed. We believe that this combination of quantitative and qualitative feedback provides a more holistic view of the perceived quality of the generated narratives. To control for potential sources of bias in our results, we collected basic demographic information from our evaluators. This data will allow us to examine any potential correlations between demographic factors and evaluation results.

IV. Experiments

A. Experimental Setup

Our experiment setup began with the collection of our dataset from [3]myanimeList.net , a comprehensive database of anime titles and synopsis. We utilized their open-source API to gather our dataset, which contained 23,044 rows of data up to the date 7th June, 2023. Each row consisted of an anime title and its corresponding synopsis. The raw data required cleaning and preprocessing to ensure its suitability for training our models. We used Python and its powerful libraries, such as NumPy and pandas, to clean the data. The cleaning process involved several steps:

- Removing Source Text: We removed any text that was not part of the actual synopsis, such as source information. This was achieved by partitioning the text at the phrases *'(Source'* and *'[Written'*, and retaining only the part before these phrases.
- Filtering Synopsis Length: We filtered out synopses that were either too short or too long. We kept only those synopses that contained between 30 and 300 words to ensure a consistent length across our dataset.
- Removing Japanese Characters and Symbols: We removed any non-ASCII characters, including Japanese characters and various symbols. This was done using regular expressions, which allowed us to efficiently search and replace these unwanted characters.

After cleaning, our dataset was ready for model training. We used the open-source GPT-2 model available from [4]Hugging Face, a leading provider of transformer-based models.

The training process was guided by a hyperparameter tuning strategy. We used the ParameterSampler from the sklearn library to sample from a predefined distribution of hyperparameters, including learning rate, batch size, epochs, and

[2]https://forms.gle/H6QTyrQpyi22UQUVA
[3]https://myanimelist.net/
[4]https://huggingface.co/

weight decay. We performed 20 iterations of this sampling process. For each set of hyperparameters, we trained the model using the AdamW optimizer and a linear learning rate scheduler with warmup. The model was trained on a CUDA-enabled GPU for efficient computation. The training process was repeated for the number of epochs specified by the current set of hyperparameters.

After each training cycle, we compared the loss to the best loss achieved so far. If the current loss was lower, we updated the best loss and saved the model's state. Upon completion of the hyperparameter tuning process, the best performing set of parameters was found to be a learning rate (lr) of 0.0001, a batch size of 10, a total of 5 training epochs, and a weight decay of 0.003. These parameters yielded the lowest loss during the training process, indicating that they were the most effective at optimizing the model's performance on our specific dataset. This optimal set of hyperparameters was then used to train the final version of our model, ensuring that we achieved the best possible performance in generating creative and engaging narratives.

In addition to our experiments with the GPT-2 model, we also conducted experiments using OpenAI's GPT-3 model. GPT-3, a state-of-the-art language model, offers significant improvements in text generation capabilities over its predecessor, GPT-2. Due to the limited access nature of GPT-3, we were unable to train the model from scratch on our dataset. However, OpenAI does provide the option to fine-tune the GPT-3 model. Fine-tuning involves training the model on our specific dataset, starting from the pre-trained model weights. This allows the model to adapt to the specific characteristics of our dataset while retaining the general language understanding capabilities it learned during its initial training. We fine-tuned the GPT-3 model on our cleaned and preprocessed dataset from MyAnimeList.net. The fine-tuning process was similar to the training process we used for GPT-2, with the goal of optimizing the model's performance on our specific task of generating creative and engaging narratives. The fine-tuned GPT-3 model was then used to generate new stories, which were evaluated using the same automatic evaluation metrics and human evaluation process as in our GPT-2 experiments.

After fine-tuning both the GPT-2 and GPT-3 models with the optimal hyperparameters, we proceeded to the text generation phase. Each model, set to evaluation mode, used parameters such as temperature, top_p, and max_length to control the diversity and length of the generated narratives. We evaluated the stories generated by both models using BERTScore and Self-BLEU metrics. BERTScore measures the semantic similarity of the generated text to a reference text, while Self-BLEU assesses the diversity of the generated text. Upon completion of the automatic evaluation, we selected the stories with the highest BERTScore and Self-BLEU from both models. These stories were deemed the best in terms of semantic similarity and originality, respectively. Finally, these top-performing stories, generated by both GPT-2 and GPT-3, were subjected to human evaluation.

In addition to the fine-tuned models, we also employed zero-shot text generation using GPT-2, GPT-3, and GPT-4. These models, without any task-specific fine-tuning, generated stories based on the same parameters of temperature, top_p, and max _length as the fine-tuned models. The stories generated by these zero-shot models were also evaluated using the BERTScore and Self-BLEU metrics to assess their semantic similarity to a reference text and the diversity of the generated text. The top-performing stories, generated by the zero-shot models, were selected based on these metrics. Finally, just like the stories from the fine-tuned models, these top-performing stories generated by zero-shot GPT-2, GPT-3, and GPT-4 were also subjected to human evaluation.

B. Results and Human Evaluation

In our study, we engaged both fine-tuned and zero-shot versions of the GPT-2, GPT-3, and GPT-4 models in narrative generation. Subsequently, these narratives were evaluated via automated measures (BERTScore and Self-BLEU) and human assessments, all of which were normalized to an identical scale ranging from 1 to 10 for effortless comparison.

To standardize the BERTScore, which naturally falls between 0 and 1 with 1 denoting perfect semantic correlation, we multiplied each individual score by 10. This method successfully reoriented the score range to 0-10, thereby allowing it to be compared directly with the other metrics.

As for the Self-BLEU score, the process of standardization was slightly more intricate. Self-BLEU also inherently resides between 0 and 1, but with 0 indicating superior quality. To rectify this inverted scale, we subtracted each original Self-BLEU score from 1, effectively reversing the orientation. This ensured that higher scores now signified better quality, harmonizing it with the other metrics. These scores were then multiplied by 10 to match the 1-10 range.

With respect to human scores, since evaluators were already instructed to rate the narratives on a 1 to 10 scale, no transformation was required.

Standardizing these scores facilitated a more intuitive comparison of the model performances, the impact of fine-tuning, and the correlation between automated metrics and human evaluations. To standardize the BERTScore and Self-BLEU scores, we utilized the following transformations:

For BERTScore, originally ranging between 0 and 1, we used a simple scaling transformation:

$$BERTScore_{standardized} = BERTScore_{original} * 10 + 1 \quad (1)$$

This transformed the BERTScore range from [0,1] to [1,10] for straightforward comparison with the human scores.

For Self-BLEU, also originally ranging from 0 to 1 but with 0 indicating superior quality, we applied a two-step process:

$$SelfBLEU_{intermediate} = 1 - SelfBLEU_{original} \quad (2)$$

$$SelfBLEU_{standardized} = SelfBLEU_{intermediate} * 10 \quad (3)$$

In the first step, we subtracted each original Self-BLEU score from 1 to reverse the scale orientation. In the second step, we multiplied the intermediate scores by 10 to bring them into the 1-10 range.

As the human evaluators rated the stories on a scale of 1 to 10, the human scores didn't require any transformations.

These standardization equations facilitated a more direct comparison between the different evaluation metrics and aligned with standard practices in the field

TABLE I
AVERAGE HUMAN SCORE, SELF-BLEU, AND BERTSCORE FOR EACH STORY BEFORE NORMALIZATION.

Story Tag	Avg. Human Score	s-bleu	b-score
GPT2_FT_1	6.28125	0.008541	0.818203
GPT2_FT_2	6.28125	0.008592	0.815351
GPT3_FT_1	7.375	0.008497	0.818651
GPT3_FT_2	8.0625	0.008037	0.818651
GPT2_Z	3.6875	0.010791	0.788979
GPT3_Z	5.78125	0.007966	0.802239
GPT4_Z	7.03215	0.008271	0.814587

Note: FT indicates 'Fine Tuned' and Z indicates 'Zero Shot'.

TABLE II
NORMALIZED AVERAGE HUMAN SCORE, SELF-BLEU, AND BERTSCORE FOR EACH STORY

Story Tag	Avg. Human Score	s-bleu_norm	b-score_norm
GPT2_FT_1	6.28125	9.923131	8.36382406949997
GPT2_FT_2	6.28125	9.922672	8.33816230297088
GPT3_FT_1	7.375	9.923527	8.36786025762558
GPT3_FT_2	8.0625	9.927667	8.24244391918182
GPT2_Z	3.6875	9.902881	8.10081470012664
GPT3_Z	5.78125	9.928306	8.22015422582626
GPT4_Z	7.03215	9.925561	8.33128190040588

Upon examining the performances of the fine-tuned models in contrast to their zero-shot counterparts, it was evident that fine-tuning considerably enhanced the narrative quality.

The fine-tuned GPT-2 model garnered an average human score of 6.28, contrasted with a significantly lower score of 3.68 for the zero-shot variant. This improvement in human rating corresponded with an increase in BERTScore from 8.10 for the zero-shot model to approximately 8.35 for the fine-tuned model. The Self-BLEU scores also slightly improved, rising from 9.90 to nearly 9.92.

Similarly, the fine-tuned GPT-3 model outstripped the zero-shot GPT-3 in terms of human evaluation, with an average score of 7.71 compared to 5.78. This superior performance was again reflected in the BERTScore, with the fine-tuned model achieving a score of 8.30 as opposed to 8.22 for the zero-shot version. The Self-BLEU scores for both models were approximately equal at 9.93.

Interestingly, although we lacked a fine-tuned version of GPT-4, the zero-shot GPT-4 model managed to produce narratives that were competitive with the fine-tuned models. This was reflected in a robust average human rating of 7.03, a BERTScore of 8.33, and a Self-BLEU score of 9.92.

To add depth to these results, we subjected the generated narratives to a human evaluation process. Evaluators were given anonymized, randomized narratives and asked to rate each one on a scale of 1 to 10. The ratings were then averaged to give a final score for each narrative.

V. DISCUSSION

The outcomes of our experiments provide several significant insights. Primarily, they reaffirm the prevalent understanding that fine-tuning significantly enhances the text generation capabilities of language models. This enhancement was evidently seen in the superior performance of the fine-tuned GPT-2 and GPT-3 models over their zero-shot counterparts, as per both automated and human evaluations.

Interestingly, our results also demonstrate that the zero-shot GPT-4 model generated narratives of a quality comparable to that of the fine-tuned models. This suggests that as language models evolve and become more sophisticated, the marginal utility of fine-tuning may decrease. This has important implications, especially in contexts where fine-tuning is not feasible due to computational or resource constraints.

Our results reveal a strong correlation between human evaluation scores and automatic metrics (BERTScore and Self-BLEU), validating the latter's utility in preliminary automatic text evaluations. While the ultimate assessment of narrative quality is best made by human readers, these automated metrics are invaluable in guiding model development and fine-tuning, enabling more objective and scalable evaluations.

However, it's important to recognize that while these metrics can identify trends and facilitate model comparisons, they have their limitations. For instance, high Self-BLEU scores can sometimes indicate a lack of diversity in the generated text. Similarly, while BERTScore is a good indicator of semantic similarity, it may fail to capture more nuanced aspects of narrative quality, such as coherence and reader engagement.

A deeper analysis of the correlation between human scores and automatic metrics revealed a more nuanced picture. While a strong correlation was observed for the narratives evaluated as being of lesser quality, this correlation was less evident for the higher-scoring narratives. The automatic metrics did not seem to capture the depth and detail appreciated by human evaluators in the higher-quality narratives, indicating that the metrics' sensitivity to narrative quality diminishes as the quality increases. This underscores the need for a more refined automatic evaluation method for high-quality text, and strengthens the argument for a combined evaluation approach that utilizes both automatic metrics and human evaluations.

Overall, while our results offer valuable insights into the narrative generation capabilities of GPT-2, GPT-3, and GPT-4 models, they also emphasize the necessity for a comprehensive evaluation approach that merges both automatic metrics and human evaluations.

VI. Conclusion

Our study explores the narrative generation capabilities of transformer-based language models, specifically GPT-2, GPT-3, and GPT-4. Through a series of experiments involving both fine-tuned and zero-shot models, we conducted an in-depth examination of the quality of the generated narratives. Utilizing a combined evaluation approach of automatic metrics (BERTScore and Self-BLEU) and human evaluations, we attempted to paint a holistic picture of the models' performance.

The results reemphasize the value of fine-tuning in enhancing the text generation abilities of language models, as demonstrated by the superior performance of the fine-tuned GPT-2 and GPT-3 models. In a striking observation, the zero-shot GPT-4 model was able to generate narratives comparable to the fine-tuned models, indicating a diminishing necessity for fine-tuning as language models become increasingly advanced.

We identified a strong correlation between automatic metrics and human evaluations for narratives of lower quality, demonstrating the usefulness of such metrics in early evaluation stages. However, this correlation weakened for high-quality narratives, suggesting the need for more refined automatic evaluation methods that can capture depth and detail in exceptional narratives.

Ultimately, our study underscores the necessity for a comprehensive, multi-faceted evaluation approach. While automatic metrics provide valuable insights and offer scalability, human evaluations remain crucial in capturing more nuanced aspects of narrative quality. As we move forward in the field of automated text generation, this combination of methodologies will be integral to the development of more advanced and nuanced language models.

VII. Future Work

The conclusions drawn from this study have provided valuable insights and open up several promising avenues for future exploration. Here are a few key areas of interest for our future work:

- **Exploring Other Language Models and Generation Strategies:** As the field of AI continues to evolve at a rapid pace, newer and more advanced language models are being developed. We aim to incorporate these models in our experiments, both to compare their performance with those we have already studied and to identify any novel attributes they may offer. In addition, we intend to explore different text generation strategies and assess their impact on the quality and diversity of the generated narratives.
- **Developing Improved Evaluation Metrics:** While BERTScore and Self-BLEU (bilingual evaluation understudy) have proven to be useful tools in our study, they each have their limitations. In future, we would like to explore and potentially develop other automatic evaluation metrics that are capable of capturing more nuanced aspects of story quality such as coherence, engagement, and plot development.
- **Conducting More Detailed Human Evaluations:** Our current approach uses an aggregate human evaluation score to provide a broad measure of narrative quality. Moving forward, we could further refine this process by asking evaluators to rate the stories on specific aspects such as creativity, coherence, or character development, providing a more detailed assessment of narrative quality.
- **Using a Diverse Dataset:** Another intriguing avenue of research involves examining the performance of these models on diverse datasets. These datasets could represent a wide range of styles, genres, and complexity levels, offering insights into how well the models can adapt to different narrative styles and constraints.
- **Fine-tuning Larger Models:** Our study was not able to incorporate a fine-tuned GPT-4 model because of limited access to it. As such, a natural extension of this work would be to fine-tune GPT-4 (or even larger, future models) and assess the potential improvements in narrative generation capabilities.
- **Generating Prompts for Other Media:** Our exploration into narrative generation can extend beyond text. The developed models could be used to create nuanced prompts for AI image or video generators. For example, AI video generators, such as the Midjourney project, could benefit from detailed, creative prompts to generate highly specific and captivating visual content.

By investigating these methods, we hope to gain a better understanding of AI models' narrative generating capabilities. This has the potential to lead to the development of more engaging, imaginative, and human-like tales across multiple areas, potentially revolutionising the field of content creation.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 30.

[2] Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q. and Artzi, Y., 2019. Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675.

[3] Akoury, N., Wang, S., Whiting, J., Hood, S., Peng, N. and Iyyer, M., 2020. STORIUM: A Dataset and evaluation platform for machine-in-the-loop story generation. arXiv preprint arXiv:2010.01717.

[4] Guan, J., Huang, F., Zhao, Z., Zhu, X. and Huang, M., 2020. A knowledge-enhanced pretraining model for common sense story generation. Transactions of the Association for Computational Linguistics, 8, pp.93-108.

[5] Correia, A.D.S. and Colombini, E.L., 2021. Attention, please! a survey of neural attention models in deep learning. arXiv preprint arXiv:2103.16775.

[6] Bahdanau, D., Cho, K. and Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

[7] Moolayil, J., 2018. Learn Keras for Deep Neural Network A fast Track Approach to Modern Deep Learning with Python, Vancouver Canada, Chapter 6 176. ISBN-13 (pbk), pp.978-1.

[8] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M. and Davison, J., 2020, October. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations (pp. 38-45).

[9] Mao, H.H., Majumder, B.P., McAuley, J. and Cottrell, G.W., 2019. Improving neural story generation by targeted common sense grounding. arXiv preprint arXiv:1908.09451.

[10] Guan, J., Wang, Y. and Huang, M., 2019, July. Story ending generation with incremental encoding and commonsense knowledge. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 6473-6480).

[11] Zhu, Y., Lu, S., Zheng, L., Guo, J., Zhang, W., Wang, J. and Yu, Y., 2018, June. Texygen: A benchmarking platform for text generation models. In The 41st international ACM SIGIR conference on research and development in information retrieval (pp. 1097-1100).

[12] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I., 2019. Language models are unsupervised multitask learners. OpenAI blog, 1(8), p.9.

[13] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S., 2020. Language models are few-shot learners. Advances in neural information processing systems, 33, pp.1877-1901.

[14] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P.J., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, 21(1), pp.5485-5551.

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

APPENDIX A
STORY GENERATED BY FINE TUNED GPT2 MODEL #1

In an alternative universe, the human race has colonized the solar system and established colonies on Mars, Earth, and the Moon. However, there is a major problem with the colonization plans: the colonists are not the only ones who are living on Earth. In response to the threat of alien life forms, a special team of colonists known as the Orbital Patrol have been sent to Earth to fight against the alien threat. One day, in an attempt to save a planet from alien invaders, they are attacked by a new type of life form called a Gargantia, which has the power to wipe out entire planets. As a result of the attacks, humanity is forced to retreat to Mars to find a way back home

APPENDIX B
STORY GENERATED BY FINE TUNED GPT2 MODEL #2

In an alternative universe, the Earth has been colonized by space pirates, who seek to conquer the universe for their own personal gain. In order to protect their planet from these alien forces, Earths only hope of survival lies in the ability to pilot giant mechas known as Titans. Titans, a race of technologically advanced humanoid beings, have the power to vanquish the forces of darkness, but are not as strong as they once were. The only way to stop them is for the Titans to become fully sentient, and this would require a great amount of energy and resources. However, this is not the only challenge these Titans face, as a new threat emerges that threatens to destroy everything they hold dear, including their very existence.

APPENDIX C
STORY GENERATED BY FINE TUNED GPT3 MODEL #1

In an alternative universe set in the distant future, a group of intergalactic space travelers embark on an epic quest to save their home planet from an unknown enemy. The group is made up of two humans, a robotic alien, and a shape-shifting alien. They must travel to different galaxies and planets in order to find the pieces of a powerful weapon that can defeat their enemy. Along the way they encounter many challenges, from battles with fierce creatures to dangerous planets full of traps. With each challenge they face, they become closer as a team and gain new abilities that help them succeed in their mission. As they draw closer to their goal, the stakes become higher and higher as they come closer to facing off against their ultimate foe. Will they be able to save their planet or will the enemy prove too powerful?

APPENDIX D
STORY GENERATED BY FINE TUNED GPT3 MODEL #1

In an alternative universe there existed a world filled with humans and magical creatures. In this world, there was a special boy by the name of Kōta who lived in a small village near a hidden forest. People from the village were advised never to enter the forest due to it being full of beasts and magical creatures that could easily cause harm or chaos. One day, Kōta decided to explore the forbidden forest against his parents' wishes. On his journey he made some surprising discoveries; magical plants that had healing properties, fairies living between trees, and even strange animals that no one had ever seen before. Despite his many encounters, Kōta felt safe while exploring the enchanted forest because he was protected by an invisible shield that seemed to always surround him. One day, he stumbled upon a hidden cave and within it contained three mysterious gems - blue, red and yellow. When Kōta touched them, he felt an incredible power surge through his body, awakening new powers that he had never felt or seen before. He soon realized that the gems granted him the ability to cast spells, control magical creatures and control any element at his will. Unbeknownst to Kōta, the gems were magic relics left by a wise wizard who witnessed potential in him from afar.

Kōta now had an incredible gift, so with it he decided to dedicate his life to protecting the people of his village and exploring more of what this magical world had to offer. Though he was still young, Kōta took on great responsibility with his newfound powers and was admired by all who knew him.

APPENDIX E
STORY GENERATED BY ZERO-SHOT GPT2 MODEL

In an alternative universe, the universe where all of the stars in the universe were created is the universe where all the galaxies were created.

We're in the universe where all of the stars in the universe were created.

But what if there's a cosmic accident in the universe that makes it impossible to create all the stars in the universe?

We're in the universe where all of the stars in the universe were created.

It's not possible to create all the stars in the universe.

But that's not the point.

APPENDIX F
STORY GENERATED BY ZERO-SHOT GPT3 MODEL

One day, a small girl named Natsume found a book called ”The Catcher in the Rye”. She read it and was charmed by its content. Then she found a book called ”The Greater Good” and learned about the different ways that good can befall someone over time. In both universes, the world is evenly balanced between those who are greater than others and those who are less so. One day, the small girl Natsume is greater than most, but due to her actions, she has caused a great rift in the balance of the world. A greater good has been called for, but Natsume is not afraid to speak out of place.

APPENDIX G
STORY GENERATED BY ZERO-SHOT GPT4 MODEL

In an alternative universe, humans coexisting with magical creatures had become the norm. Among these humans, a young boy named Hiroki found himself blessed - or cursed - with the ability to see the future. Born into a simple farming family, Hiroki had lived a sheltered life, plagued with surreal dreams and haunting visions. He saw entire kingdoms fall, heroes rise, and magical creatures fighting cataclysmic battles. However, despite his ability, Hiroki felt powerless until the day a crystal-eyed wyvern named Vira landed in his field.

Vira revealed that Hiroki’s visions weren’t mere nightmares, but prophecies. Hiroki was destined to prevent his visions from becoming reality, averting the disastrous future he foresaw. Despite feeling anxious about his newfound responsibilities, Hiroki was determined to change destiny. With Vira as his trusty guide and companion, they journeyed across their universe filled with dangers, all banking on Hiroki’s hidden strength and threatened by his daunting visions.