Interactive Story Generation

Mahisha Ramesh
IIITD
MT23121, India
mahisha23121@iiitd.ac.in

Shaina Mehta
IIITD
MT23139, India
shaina23139@iiitd.ac.in

Mahima Chopra
IIITD
2021398, India
mahima21398@iiitd.ac.in

Janesh Kapoor
IIITD
2021466, India
janesh21466@iiitd.ac.in

Shivam Dwivedi
IIITD
2021352, India
shivam21352@iiitd.ac.in

ABSTRACT

Storytelling [1] has been one of the most essential parts of human culture for many years. People tell stories to others to share their experiences, beliefs, and values via paintings, carvings, movies, podcasts, etc. Technological advancements and increased Artificial Intelligence have expanded our ability to tell stories. AI-based storytelling has become popular, but producing consistent, coherent, and engaging narratives takes a lot of work. This encompasses issues such as maintaining logical plot progression, developing well-rounded characters, and ensuring that the story's pacing remains appropriate. The AI may struggle with generating stories that flow naturally and captivate the audience, leading to disjointed or unsatisfying narrative experiences. Additionally, balancing creativity with coherence presents a complex problem, as the AI must innovate within established storytelling frameworks while still adhering to logical constraints. This project aims to build a platform for interactive story generation by finetuning Large Language Models, such as GPT-3.5 turbo,google gemma models based on different themes. It also focuses on developing stories, ensuring consistency, coherency, and engaging narratives.

KEYWORDS

Storytelling, Narrative, Character Development, Plot Progression, Coherency, Engagement, AI, Interactive Story Generation

ACM Reference Format:
Mahisha Ramesh, Shaina Mehta, Mahima Chopra, Janesh Kapoor, and Shivam Dwivedi. 2018. Interactive Story Generation. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 MOTIVATION

With this project, we want to present the utility of Large Language Models for generating narratives. The idea behind this project is that engaging storytelling is fundamental to capturing and retaining users' attention, fostering immersion, and eliciting emotional responses. Without compelling narratives, users are less likely to interact with the AI generator, leading to diminished satisfaction and decreased utility of the tool. Moreover, it is crucial to solve this issue since it is necessary to expand the potential of Generative AI in several fields, such as therapy, education, and entertainment. It also has the potential to increase the explainability of the AI models, especially Large Language Models, to help researchers enhance the AI models for various applications in future [2].

2 LITERATURE SURVEY

Several researchers have done various research work in AI-based story generation. Earlier works were based on logic, formal grammar and statistical machine-learning models. Nowadays, Deep Learning plays a crucial role in generating engaging narratives with high consistency and coherence. Fan et al. [3] have developed the dataset for story generation that is Writing Prompts, which has approximately 300K stores along with its prompts and trained fusion of Conv Seq2seq + self-attention-based model and achieved the perplexity of 36.08 and 36.56 on validation and testing set respectively and performed well on human evaluation test. Khan et al. [4] have developed a keyword-based story generation model by finetuning the GPT 2 model on the private dataset they created. They achieved the BLEU score of about 0.704, averaging over ten genres. Yao et al. [5] proposed a framework for generating the stories called Play-and-Write, which plans the storyline and generates the stories. It works in two static and dynamic schemas, which have performed better than the baselines in objective and subjective evaluation. Pradyumna et al. [6] used the Reward Shaping technique under Reinforcement Learning to generate stories. They trained the models on the CMU movies dataset and achieved lower perplexity values (7.61 and 5.73 for DRL Clustered and DRL Unrestricted with the goal 'admire') than the baseline model. They performed better in human evaluation. Ammanabrolu et al. [7] developed an ensemble-based framework for generating stories based on a combination of the Retrieve and Edit model, Sentence Templating, Monte-Carlo Beam Search and Finite State Constraint Search algorithms. They achieved the lowest perplexity and highest BLEU scores of 70.179 and 0.0481, respectively, compared to all four models applied individually. Kong et al. [8] developed the story generator model, which plans the stylised keywords and then generates the stories based on these keywords. This model achieved the LSC and SSC scores of 0.474 and 0.371 for emotion-driven style and 0.309 and 0.293, respectively, for event-driven style, higher than the baseline models. Mathews R.F. et al. [9] have developed the procedural story generation model using a handcrafted event network and a dynamic artificial social network created for each new story. SeokKyoo Kim et al. [10] have developed a story generation algorithm using a Constraint Narrative structure implemented in Storytelling Markup Language.

3 NOVELTY

The novelty is introduced to the dataset, and the system introduces the multimodality aspect. Unlike other story generation datasets, this dataset is collected from various websites, taking care of repeating the stories in the corpus and not including inappropriate stories as seen in other corpora. The second novelty is introduced on the website, where we have deployed the trained model by introducing the story recitation feature that will recite the story using OpenAI's API.

4 DATASET

We have scrapped about 356 stories from websites like Medium, Reddit, Project Gutenberg, etc., based on six themes: adventure, horror, humour, mystery, romantic and sci-fi. However, the horror stories have been taken directly from the dataset used during baseline results. Some stories that exceed the word limit of 500 words are summarised using chat GPT 3.5. Based on the available computing resources, we worked with 100 adventure stories, 100 sci-fi stories and 100 horror stories. For each theme, nine stories are used for training, three for validation, and three for model testing. Figures 1, 2, and 3 show the maximum token length of the stories of each split per theme.

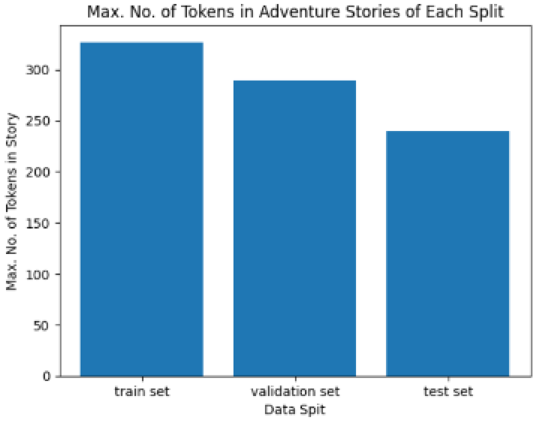


Figure 1: Maximum Number of Tokens of Adventure Stories of Each Split

In adventure stories, the maximum number of tokens in the training, validation, and test sets are 327, 290, and 239, respectively. In horror stories, the maximum number of tokens in the training, validation, and test sets are 507, 803, and 348, respectively. In adventure stories, the maximum number of tokens in the training, validation, and test sets are 307, 310, and 257, respectively.

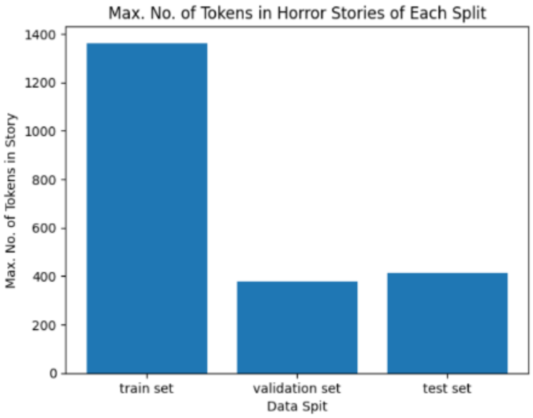


Figure 2: Maximum Number of Tokens of Horror Stories of Each Split

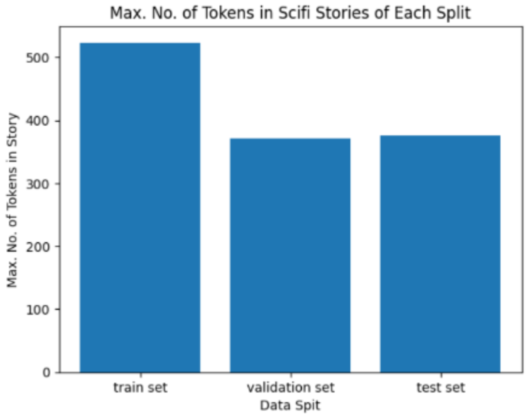


Figure 3: Maximum Number of Tokens of Scifi Stories of Each Split

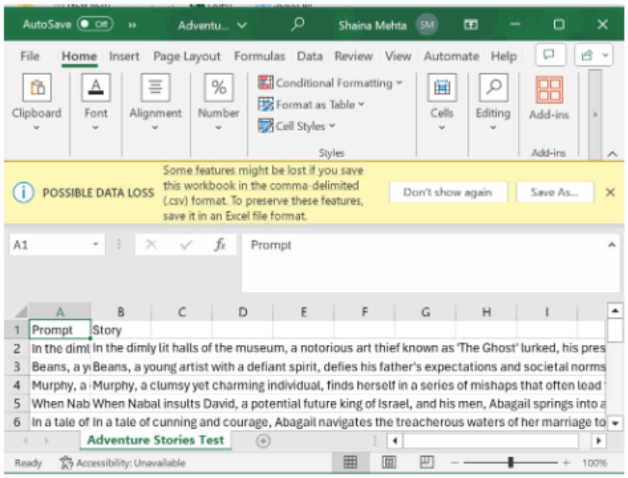
5 METHODOLOGY

5.1 Data Pre-processing and Finetuning of Gemma2B Instruct Model

The following steps are followed for data pre-processing and finetuning of the Gemma2B Instruct model:

- The prompts are created by taking the story's beginning, and each story and its corresponding prompt is saved in the .csv files, as shown in Figure 4.
- Then, the prompts and the stories are combined in a conversational format, as suggested by [11].
- The Gemma2B Instruct model is loaded from the hugging face library, and then it is quantised using LoRA and QLoRA so that it will not occupy too much space on the GPU, as suggested by [11].
- Then, the model is finetuned for each theme to 4 epochs, with a learning rate of 0.0001 using a paged-adamw 8-bit

optimiser, and the loss curves are plotted. Here, we have used 75 data samples for training, 10 for validation and 15 for testing.
- The model is merged back to the original precision value for evaluation purposes, as suggested by [11].



POSSIBLE DATA LOSS Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.

A1 Prompt

	A	B	C	D	E	F	G	H	I
1	Prompt	Story							
2	In the dim	In the dimly lit halls of the museum, a notorious art thief known as 'The Ghost' lurked, his pres							
3	Beans, a y	Beans, a young artist with a defiant spirit, defies his father's expectations and societal norms							
4	Murphy, a	Murphy, a clumsy yet charming individual, finds herself in a series of mishaps that often lead							
5	When Nab	When Nabal insults David, a potential future king of Israel, and his men, Abagail springs into a							
6	In a tale of	In a tale of cunning and courage, Abagail navigates the treacherous waters of her marriage to							

Figure 4: Sample Dataset for Finetuning Gemma2B Instruct Model

5.2 Data Pre-processing and Finetuning of GPT 3.5 Turbo Model

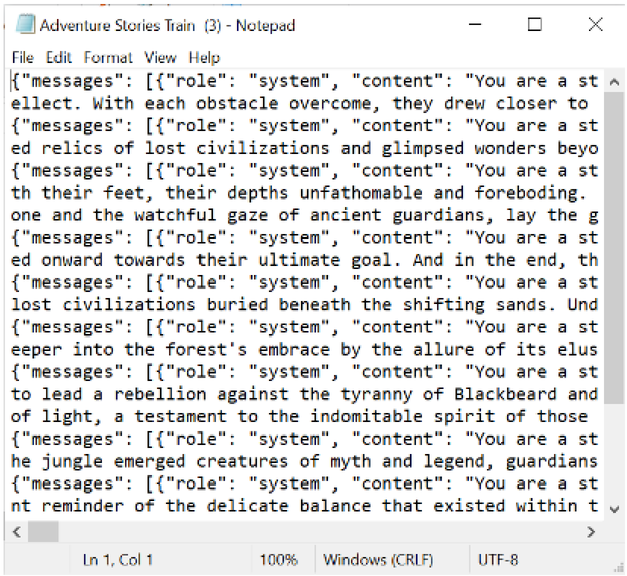
The following steps are followed for data pre-processing and finetuning of the GPT 3.5 Turbo model: Generating the prompts of the stories in the form of the story's title using ChatGPT 3.5 and format it in the conversational format suggested by [12] as shown in figure 5, and saving it into .jsonl format. Loading the GPT 3.5 model using the OpenAI API key and finetuning it for each theme for 46 epochs and plotting its loss curves and further evaluation is performed. Here, we have used 9 data samples for training, 3 for validation and 3 for testing.

6 RESULTS

This section discusses the results of the models. All the models are trained using Python 3.10 Programming language using Google Colab. The plots for training loss for Gemma 2B Instruct and GPT 3.5 Turbo model trained on Adventure, Horror and Scifi stories separately, BLEU and Mean Perplexity scores on the testing set of each theme trained on Gemma2B Instruct model and their description is given in the section 4.1 and 4.2.

6.1 Gemma2B Instruct Model

Figures 6 show the training and validation loss curves of the Gemma2B Instruct model trained on adventure, horror and sci-fi themes, respectively. From them, one can infer that the model is learning properly, but the learning process needs to be completed due to the lack of computation resources.



```
{"messages": [{"role": "system", "content": "You are a st
ellect. With each obstacle overcome, they drew closer to
{"messages": [{"role": "system", "content": "You are a st
ed relics of lost civilizations and glimpsed wonders beyo
{"messages": [{"role": "system", "content": "You are a st
th their feet, their depths unfathomable and foreboding.
one and the watchful gaze of ancient guardians, lay the g
{"messages": [{"role": "system", "content": "You are a st
ed onward towards their ultimate goal. And in the end, th
{"messages": [{"role": "system", "content": "You are a st
lost civilizations buried beneath the shifting sands. Und
{"messages": [{"role": "system", "content": "You are a st
eeper into the forest's embrace by the allure of its elus
{"messages": [{"role": "system", "content": "You are a st
to lead a rebellion against the tyranny of Blackbeard and
of light, a testament to the indomitable spirit of those
{"messages": [{"role": "system", "content": "You are a st
he jungle emerged creatures of myth and legend, guardians
{"messages": [{"role": "system", "content": "You are a st
nt reminder of the delicate balance that existed within t
```

Figure 5: Sample Dataset for Finetuning GPT 3.5 Turbo Model

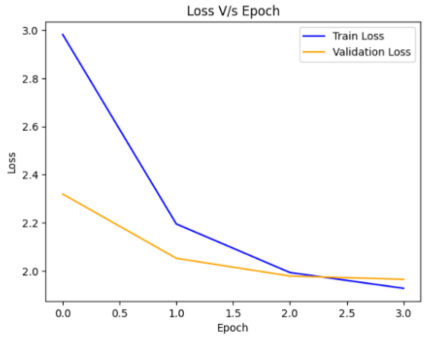


Table 1 shows the BLEU and Mean Perplexity scores of the Gemma2B Instruct model trained separately on Adventure, Sci-Fi and Horror themes. From this, one can infer that horror story generator model is far better than the adventure and sci-fi story generator models concerning the perplexity score, and the adventure story generator is far better than sci-fi and horror story generation models in terms of BLEU Score 2, 3 and 4. However, sci-fi story generator model is better than the horror and adventure story generator model in terms of BLEU Score 1 but all of the model performances are not upto the mark.

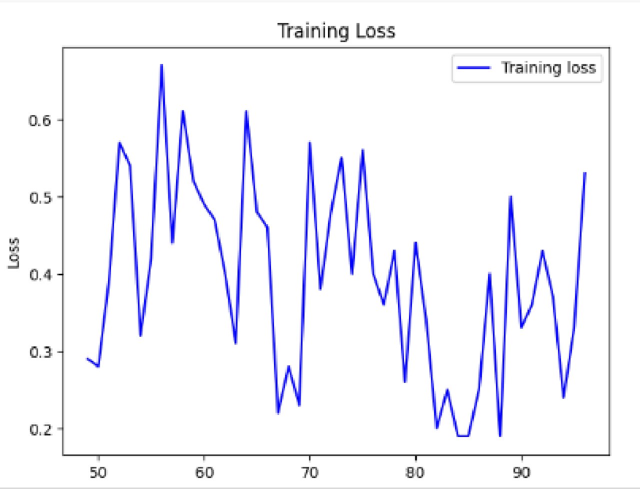
On human evaluation, it has been observed that the generated stories could be more up to the mark; some of them are not aligned to the theme and some have repetition of words and phases. However, the results are far better than those of the baseline.

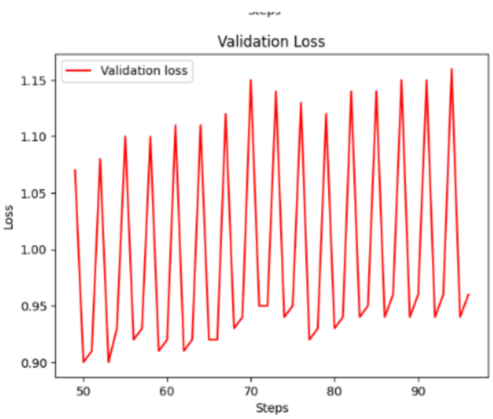
Table 1: BLEU and Mean Perplexity Scores of Stories of Each Theme

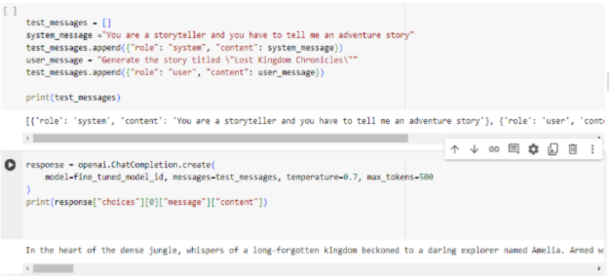
Theme	BLEU 1 Score	BLEU 2 Score	BLEU 3 Score	BLEU 4 S
Adventure	0.317	0.173	0.109	0.109
Horror	0.310	0.160	0.098	0.073
Sci-Fi	0.348	0.187	0.121	0.092

6.2 GPT 3.5 Turbo Model

Figures 12, 13, 15 show the training and validation loss curves of the GPT 3.5 Turbo Model trained on adventure, sci-fi and horror themes, respectively. From these curves, one can infer that training loss and validation loss fluctuate. This means that the model is overfitting. Regarding human evaluation, the generated stories are better than the Gemma 2B Instruct model, but the narratives could be better.







```
test_messages = []
system_message ="You are a storyteller and you have to tell me an adventure story"
test_messages.append({"role": "system", "content": system_message})
user_message = "Generate the story titled \"Lost Kingdom Chronicles\""
test_messages.append({"role": "user", "content": user_message})

print(test_messages)
```

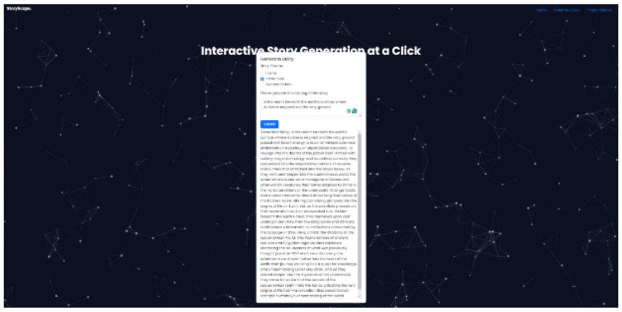
```
[{'role': 'system', 'content': 'You are a storyteller and you have to tell me an adventure story'}, {'role': 'user', 'cont
```

```
response = openai.ChatCompletion.create(
    model=fine_tuned_model_id, messages=test_messages, temperature=0.7, max_tokens=500
)
print(response["choices"][0]["message"]["content"])
```

```
In the heart of the dense jungle, whispers of a long-forgotten kingdom beckoned to a daring explorer named Amelia. Armed w
```

7 THE WEBSITE

The website has been created using JavaScript on the front end and Python on the back end. The website is running on the Microsoft Azure Server. The website has the home page, the first page of the website, the report page, which has the project report, and the storyteller page, in which the user will give the story's beginning, select the story's theme and click on the submit button. The model will generate the story and return it to the front end.



8 CONCLUSION AND FUTURE WORK

We have fine-tuned the GPT-3.5 and Gemma2B Instruct models on stories based on three themes, and they are giving better performance than baseline results but could have done better. We planned to increase the dataset by taking the actual stories for many words, training the same models, and deploying them.

REFERENCES

(1) AIContentfy. (2023, November 6). The art of AI-generated storytelling. AIContentfy. Retrieved January 31, 2024, from https://aicontentfy.com/en/blog/art-of-ai-generated-storytelling#:~:text=One%20potential%20future%20for%20AI,may%20not%20have%20considered%20before.

(2) Riedl, M. (2021, January 4). An Introduction to AI Story Generation. Medium. Retrieved February 1, 2024, from https://mark-riedl.medium.com/an-introduction-to-ai-story-generation-7f99a450f615.

(3) Fan, A., Lewis, M., & Dauphin, Y. (2018). Hierarchical neural story generation. arXiv preprint arXiv:1805.04833.

(4) Khan, L. P., Gupta, V., Bedi, S., & Singhal, A. (2023, April). StoryGenAI: An Automatic Genre-Keyword Based Story Generation. In 2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES) (pp. 955-960). IEEE.

(5) Yao, L., Peng, N., Weischedel, R., Knight, K., Zhao, D., & Yan, R. (2019, July). Plan-and-write: Towards better automatic storytelling. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 7378-7385).

(6) Pradyumna, T., Murtaza, D., Lara, J. M., Mehta, A., & Harrison, B. (2019). Controllable neural story plot generation via reward shaping. In Proc. Int. Joint Conf. Artificial Intelligence (pp. 5982-5988).

(7) Ammanabrolu, P., Tien, E., Cheung, W., Luo, Z., Ma, W., Martin, L., & Riedl, M. (2019, August). Guided neural language generation for automated storytelling. In Proceedings of the Second Workshop on Storytelling (pp. 46-55).

(8) Kong, X., Huang, J., Tung, Z., Guan, J., & Huang, M. (2021). Stylised story generation with style-guided planning. arXiv preprint arXiv:2105.08625.

(9) Jain, P., Agrawal, P., Mishra, A., Sukhwani, M., Laha, A., & Sankaranarayanan, K. (2017). Story generation from a sequence of independent short descriptions. arXiv preprint arXiv:1707.05501.

(10) Mendonça, M. R., & Ziviani, A. (2018). Network-based procedural story generation. Computers in Entertainment (CIE), 16(3), 1-18.

(11) Kim, S., Moon, S., Han, S., & Chan, J. (2011). Programming the story: Interactive storytelling system. Informatica, 35(2).

(12) Adithya S K. (2024, February 22). A Beginner's Guide to Fine-Tuning Gemma. Medium. Retrieved March 30, 2024, from https://adithyask.medium.com/a-beginners-guide-to-fine-tuning-gemma-0444d46d821c.

(13) Youssef Hosni (2024, February 22). 14 Free Large Language Models Fine-Tuning Notebooks. Medium. Retrieved March 27, 2024, from https://levelup.gitconnected.com/14-free-large-language-models-fine-tuning-notebooks-532055717cb7.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009