Improving Consistency for Story Generation in LLaMA 3.1

Rijul Saigal, Hari Shankar, Nikhil Raj Sriram

September 2024

1 Introduction

1.1 Story Generation

AI Generated Stories generally involve the usage of Language Models, commonly LLMs, that have been previously trained on large corpora of text. Models can generate content based on prompts, producing anything from short anecdotes to stories spanning multiple chapters.

One of the issues noted in the current state of AI story generation is that the output often appears logically inconsistent with itself. Inconsistent elements hinder the reader's engagement, and may often confuse the reader. Characters might exhibit unpredictable behaviors, the story-line might lack coherent development, or key plot points may contradict one another, to name a few potential issues.

1.2 Project Objectives

For our project, we shall limit our scope to Flash Fiction, or works of Fiction that range between 300 to 1000 words. Additionally, we shall only consider the LLaMA 3.1 8B model, as available to us on Huggingface, without any additional fine-tuning. We aim to build a system that can generate a piece when provided a prompt by the user, while improving on the following dimensions, with respect to consistency:

- Character Consistency: Current systems often generate characters that change their personalities, motives, or actions without sufficient justification, leading to a breakdown in believability.
- Plot Consistency: The flow of events in stories generated by current systems can sometimes diverge or contradict earlier developments, leading to confusion in plot progression.
- Worldbuilding Consistency: Elements of the story's world, such as rules, environments, and technologies, can shift unexpectedly, undermining immersion and the reader's suspension of disbelief.

- Tone and Style Consistency: The overall tone or style of the story may shift abruptly, from serious to comedic or from formal to casual, creating an inconsistent reader experience.

2 Experiments

On a broad level, we plan to improve consistency through pre-writing; we draw inspiration from the way human writers often plan and structure their stories before actually starting the writing process. Pre-writing will serve as a guiding framework for the AI, ensuring that the generated narrative follows a more consistent and coherent trajectory.

We aim to clearly define essential elements of a story, such as characters, plot points, themes, settings, and tone before the actual story generation begins.

We take inspiration from work by Shao, Yijia, et al. [1] in generating and using different point-of-views to first ask questions on a given topic (here, plot, character motives, world-building, etc.), and then building an outline from this information.

We subsequently use the outline to generate our story.

2.1 Different Methods in Obtaining Outline

We check the following approaches for pre-writing, as introduced by Shao, Yijia, et al., in our domain of story generation:

- Direct Prompting
- Perspective-Guided Question Asking

Pre-writing would be required for improving characters, world-building as well as plot - and so, these approaches must be tested on all.

2.2 Outlines from Story Frames

We check whether story generation can be improved by the usage of different frames to stories;

- No Guidance in Forming Outline
- Outline to follow the Story Mountain narrative structure (of Opening, Build-Up, Problem, Resolution, End)
- Outline to follow Dan Harmon's Story Circle narrative structure [2]

3 Procedure

The model selected by us was the LLaMA 3.1 8B variant; however, due to our computational constraints, it was not possible to use the model as is, and instead had to be quantized to an 8-bit form. We accessed this model from HuggingFace. Model inferences were made using the llama-cpp-python library, whose repo can be found here.

3.1 Direct Prompting

The following questions were used, and the entire chat history is used in generating the final story;

1. Here's the idea: "" What genre(s) could this story fall into? Answer in 1 sentence.
2. What kind of message do you want the story to have? Answer in 1 sentence.
3. Are there any sub-themes you want to incorporate? Answer in 1 sentence.
4. Summarize the plot of the story for me in a single line.
5. In what time period does your story take place? Answer in 1 sentence.
6. What's the primary setting or location? Answer in 1 sentence.
7. Are there any unique aspects of this world that affect the story? Answer in 1 sentence.
8. Are there any social, political, or technological elements that are important? Answer in 1 sentence.
9. Who is your protagonist? What do they want?
10. What obstacles stand in their way?
11. Who are the other key characters?
12. What are the relationships between characters?
13. What motivates each character's actions?
14. What flaws or quirks make your characters interesting?
15. What inciting incident sets the story in motion? Answer in 1 sentence.
16. What are the key events or turning points?
17. What is the main conflict (internal, external, or both)?
18. How will the tension escalate?

19. What's at stake for the characters?

20. How might the story end? What changes for the characters?

21. Which scene or moment will you focus on?

22. What context can you imply rather than state directly?

23. Which details are essential to include?

24. What sensory elements can quickly establish the setting?

25. Make me an outline for making a flash fiction story using all of this information. This story will have a maximum word limit of 700-1000 words; so you would only be able to focus on the one scene. Having said that, break this scene up into beats for the outline. Do not plan for sudden jumps in time/location between these beats, unless absolutely necessary. Add some details about the characters, and the setting of this scene at the end. Use a maximum of 500 words.

26. Write a story of between 700-1000 words using this outline.

These questions were manually selected to simulate the process of making an outline from a basic idea.

In addition to this, a system prompt was given to the model to generally limit responses to a maximum of 5 sentences, unless stated otherwise - to stay within the context length of 4096 tokens for the Llama model.

3.2 Model for evaluating textual coherence

For the purpose of evaluating the coherence of the text generated by our model, we trained a transformer model developed in this paper [3]. We trained this model on the task of classifying a given text as having high, medium or low coherence. We generated some example short stories and modified the source code of the transformer model to test it on those short stories.

3.3 BLEU, ROUGE, BERTScore

The BLEU, ROUGE and BERTScore metrics are easily accessbile using the evaluate library; scores shall be obtained on the stories generated from our system using the Outline as reference, and using the story output as the prediction.

4 Evaluation Methodology

4.1 Reader Engagement

We evaluate the Reader Engagement through Surveys and Questionnaires, asking the respondent questions about Consistency of Characters, Plot, Worldbuilding, etc. for stories generated both by our system, and through direct prompting of the LLM.

4.2 Perplexity

We evaluate the perplexity score of the story generated by the model. This will indicate how confident the model is in its ability to generate a grammatically correct output.

5 Timeline

1. Outline proposal (September 18, 2024)
2. Pipeline to evaluate Textual Coherence of Text
3. Pipeline - Direct Prompting Approach.
4. Interim Submission (October 5, 2024)
5. Pipeline - Perspective-Guided Approach.
6. Pipeline - To follow different Narrative Structures.
7. Survey Questionnaire to be sent with genearated stories (October 25, 2024)
8. (Extra) - Swapping LLaMA 8B with other models
9. Final submission - Final Submission (November 12, 2024)

References

[1] Yijia Shao, Yucheng Jiang, Theodore A Kanell, Peter Xu, Omar Khattab, and Monica S Lam. Assisting in writing wikipedia-like articles from scratch with large language models. *arXiv preprint arXiv:2402.14207*, 2024.

[2] Dan Harmon. Story structure 101: Super basic shit, 2014.

[3] Tushar Abhishek, Daksh Rawat, Manish Gupta, and Vasudeva Varma. Transformer models for text coherence assessment. *arXiv preprint arXiv:2109.02176*, 2021.