TeLLMyStory: a LLM for stories generation aiming the reinforcement of the controllability aspect

Thomas BLUMET[a,*], Thomas HALVICK[a,] and Ethan LE TRUNG[a,***]**

[a]Students in AI PhD, University Lyon 1

Abstract. Recent advancements in Large Language Models (LLMs) have significantly improved their ability to generate creative, coherent, and contextually relevant text, particularly for storytelling. However, the issue of controllability remains a major challenge. Controllability refers to the ability of users to effectively guide the model's output through detailed prompts, such as specifying key events, character traits, and narrative tone. This paper presents the TeLLMyStory project, which focuses on enhancing the controllability of story generation using the pre-trained Mistral7B model. Although the model was not fine-tuned during this study, it was designed to address the challenges of story generation while maintaining a balance of flexibility and creativity, with the aim of developing more adaptable AI-driven storytelling systems that better meet users' specific needs.

1 Introduction

In recent years, the field of natural language processing (NLP) has made remarkable progress with the development of large language models (LLMs), enabling the generation of highly coherent, creative, and contextually relevant text. These advancements have opened up new possibilities for AI-driven storytelling, allowing models to generate narratives with impressive fluency. However, a significant challenge persists: controllability. Controllability refers to the ability of users to guide the model's output effectively, ensuring the generated story aligns with specific instructions such as tone, plot structure, character development, and style [1, 10].

While large language models have demonstrated impressive creativity, they often fail to consistently meet user expectations, particularly when it comes to more complex narrative constraints. These models may struggle to maintain coherence, prevent deviations from the intended plot, or handle nuanced storytelling elements like character voice or emotional tone [3]. As a result, achieving a balance between creativity and precise control remains an open problem.

The TeLLMyStory project aims to address these challenges by exploring how LLMs can be more effectively guided by user input, enabling the generation of stories that not only retain creativity but also adhere more closely to user-defined narrative parameters. This paper discusses the potential of improving story generation through better controllability, emphasizing how large pre-trained models like Mistral7B can be leveraged to create more adaptable, user-driven storytelling systems [4].

2 State of the Art

Recent advancements in story generation using large language models (LLMs) have demonstrated significant improvements in creativity and fluency, allowing models to generate highly coherent narratives. However, one of the most prominent challenges remains the issue of controllability, particularly when users aim to guide the model's output in specific ways. Controlling the direction of a story is difficult because LLMs, despite their impressive performance, may not always adhere to user-defined instructions, such as maintaining a consistent narrative tone, plot structure, or character development [1, 10].

Several methods have been proposed to address this challenge, with varying degrees of success. Early approaches focused on conditioning models on explicit prompts that outline key elements of the story, such as character traits, plot points, or emotional tone. These techniques helped LLMs generate stories that were more aligned with user inputs, but they still struggled with long-term coherence and maintaining narrative control over extended text [3].

In addition, reinforcement learning (RL) approaches have been explored to improve controllability by optimizing for specific goals set by the user, such as coherence, creativity, or adherence to the prompt [4, 9]. Some models, including BART and T5, have shown promise in controlling narrative direction by conditioning on specific prompts or styles, but challenges persist, especially in maintaining consistency and avoiding biases or irrelevant content [5]. These methods, while effective to some degree, often encounter limitations when scaling to more complex or longer-form stories.

Recent works have also begun exploring the integration of common sense reasoning and external knowledge sources to enhance the model's understanding of the narrative world. However, while these techniques show potential in improving the model's contextual understanding, they do not fully resolve the issue of user control. Furthermore, the challenge of ensuring that the generated text aligns with specific user expectations, particularly in long-form storytelling, remains a topic of active research [6].

Overall, despite the advancements made, controllability in LLM-driven story generation is still an open problem. Existing methods provide some level of control but are often limited by factors such as the model's inability to maintain coherence over extended narratives, the difficulty of explaining decision-making processes, and the persistence of biases in the generated content. Addressing these issues is crucial for the development of more reliable and adaptable AI-driven storytelling systems.

* Corresponding Author. Email: thomas.blumet@etu.univ-lyon1.fr

** Corresponding Author. Email: thomas.halvick@etu.univ-lyon1.fr

*** Corresponding Author. Email: ethan.le-trung@etu.univ-lyon1.fr

3 Methodology

3.1 Choice of Text Generator

The challenge of controllability in story generation begins with selecting an appropriate text generator. In this project, we focus on the Mistral7B model, a pre-trained large-scale language model with over 7 billion parameters. This model has shown strong performance in various natural language processing tasks, including text generation, and its advanced transformer architecture makes it particularly suited for producing coherent, fluent, and contextually relevant text [7]. The coding framework for the model was implemented and tested in a Colab notebook environment, inspired by a notebook reference [2], which provided the tools for evaluation process in our study.

However, even though Mistral7B has strong generative capabilities, its ability to follow specific, user-defined instructions—such as adhering to a particular narrative tone, character development, or plot structure—remains a challenge. This issue of controllability becomes even more important when the goal is to create narratives that meet specific user expectations while still maintaining the creativity and flexibility of the model.

3.2 Controllability through Prompts

To address the challenge of controlling the narrative direction, the approach focuses on carefully crafting and utilizing detailed prompts that guide the model's generation process. These prompts allow users to specify certain aspects of the story, such as key events, character traits, or emotional tone, providing some level of control over the narrative output [8]. The effectiveness of these prompts in guiding the model toward desired outcomes is central to evaluating the controllability of the generated stories.

3.3 Evaluation of Controllability

To evaluate how well the model responds to user inputs and adheres to defined narrative constraints, both quantitative and qualitative evaluation methods are employed.

3.3.1 Quantitative Metrics

Metrics such as perplexity, ROUGE, and METEOR are used to measure the fluency and consistency of the generated stories. Perplexity assesses how well the model predicts the next token in the sequence, while ROUGE and METEOR measure the similarity of the generated text to reference texts, with an emphasis on content and semantic meaning.

3.3.2 Qualitative Metrics

Human evaluation is an essential component of assessing controllability. Human assessors rate the generated stories based on various factors, including coherence, relevance to the user prompt, and creativity. By evaluating the generated text from a more subjective perspective, these assessments provide insights into how well the model adheres to specific narrative guidelines, as well as its ability to generate engaging and high-quality stories.

3.4 Evaluation of Controllability

One of the primary challenges in improving controllability is balancing the model's creativity with the ability to meet user-defined constraints. Although Mistral7B is capable of generating diverse and interesting narratives, the model sometimes deviates from the intended story direction or does not fully align with the specific requirements set by the user. The inherent unpredictability of large language models means that even with detailed prompts, the generated narratives may not always meet expectations, particularly in long-form stories where maintaining coherence and consistency is more difficult.

Additionally, the "black-box" nature of these models makes it challenging to fully understand or explain why the model chooses certain narrative directions over others. This lack of transparency can make it difficult to ensure complete control over the generated output and to provide users with more intuitive ways to guide the model's storytelling process.

Despite these challenges, this project aims to explore techniques for improving controllability, such as refining prompt structures, optimizing evaluation methods, and investigating new strategies for guiding the model's output to meet user expectations while preserving its creative flexibility.

4 Results and Discussion

4.1 Training Performance

Although we did not perform full training on the model, we focused on assessing the effectiveness of the prompts used for guiding the model's story generation. Using a selected input prompt, "The neverending day began with a beautiful sunshine and an AI robot which was seeking humans on the desert Earth," we analyzed the performance of the generated output using standard evaluation metrics such as ROUGE.

4.2 Evaluation Metrics

To evaluate the controllability and quality of the generated text, we used the ROUGE metric, which is commonly used in the text generation field to assess the overlap between n-grams in the generated text and reference text. The specific ROUGE scores obtained for the example input prompt are as follows.

Table 1. Metric values for a given prompt.

ROUGE-1	0.3401
ROUGE-2	0.0966
ROUGE-L	0.2041
ROUGE-Lsum	0.2449

4.2.1 ROUGE-1: 0.3401

The ROUGE-1 score reflects the overlap of unigrams (single words). A score of 0.3401 indicates moderate alignment in terms of vocabulary, showing that the model captured key words from the prompt, but there's still room for improvement in fully matching the reference content.

4.2.2 ROUGE-2: 0.0966

ROUGE-2 measures the overlap of bigrams (pairs of adjacent words). With a score of 0.0966, this suggests that while the model generates relevant words, it struggles to maintain coherence at the phrase level, particularly in more complex sentence structures.

4.2.3 ROUGE-L: 0.2041

ROUGE-L evaluates the longest common subsequence (LCS), indicating how well the model preserves the sequence of words. A score of 0.2041 suggests that while the model keeps some structure, it does not fully align with the narrative flow of the reference text.

4.2.4 ROUGE-Lsum: 0.2449

ROUGE-Lsum evaluates how well the model captures the overall meaning in a summarized form. This score of 0.2449 shows some alignment with the reference, but like the other scores, indicates that the model doesn't perfectly capture the essence of the intended narrative.

4.3 Interpretation of Results

The results show that while the model can generate text that is somewhat aligned with the input prompt, there are areas where improvement is needed. The relatively moderate ROUGE-1 score suggests that the model is able to produce relevant vocabulary and follow the basic instructions in terms of content. However, the low ROUGE-2 score reveals that the model struggles with generating more complex or syntactically accurate sentences, which might impact the flow of the story.

The ROUGE-L score indicates that the model captures the general sequence and structure of the story to some extent, but there are still inconsistencies in maintaining a coherent narrative. The ROUGE-Lsum score, although better than the ROUGE-L, reflects similar challenges in fully capturing the essence of the narrative as intended by the user.

Overall, these results highlight that while the model shows some potential in terms of content generation, there is still considerable room for improvement, especially in ensuring that the generated text aligns more closely with the intended narrative flow, structure, and syntactic complexity.

4.4 Challenges Encountered

The analysis of the ROUGE scores highlights some inherent challenges in controlling story generation using large language models. Specifically, while the model demonstrates reasonable performance in capturing individual words (ROUGE-1), it faces difficulties in generating more complex sequences (ROUGE-2) and in maintaining the overall structure of the narrative (ROUGE-L). These challenges point to the difficulty of balancing creativity and coherence, which is an ongoing concern in the field of AI-driven story generation.

Another challenge is ensuring that the model's creativity does not lead to deviations from the intended narrative. While the model is able to produce novel and engaging content, fine-tuning its ability to follow more precise user instructions remains a key goal for improving controllability.

5 Future Work

Future work will focus on improving the model's controllability and alignment with user-defined prompts. Key areas of development include:

- Fine-Tuning: We will fine-tune the model on specific datasets to improve its coherence, tone, and character consistency based on user instructions [4].

- Enhanced User Control: Techniques like reinforcement learning from human feedback (RLHF) could provide finer control over narrative elements such as plot and character development [8].

- Bias Mitigation: Additional work will focus on detecting and reducing biases in generated content, ensuring more inclusive and balanced narratives.

- Scalability: Optimizing the model for large-scale generation and efficient inference will be crucial for real-world applications.

6 Conclusion

This paper presents TeLLMyStory, a project aimed at enhancing the controllability of story generation using large language models. Through the evaluation of the model's performance with specific input prompts, we demonstrated the model's ability to generate relevant and coherent stories, but also highlighted areas for improvement, particularly in maintaining narrative structure and syntactic complexity.

Future work will focus on fine-tuning the model, improving user control, and addressing biases in the generated narratives. Despite the current limitations, this work lays the foundation for more adaptable and controllable AI-driven storytelling systems. By refining the model's alignment with user instructions and enhancing its ability to generate high-quality, personalized stories, we hope to contribute to the development of more engaging and reliable AI-assisted storytelling technologies.

7 Links and references

GitHub project link : https://github.com/ThomasBlumet/TeLLMyStory
Youtube Video Link : https://youtu.be/T2vVyhk1OBo

References

[1] A. Alabdulkarim, S. Li, and X. Peng. Automatic story generation: Challenges and attempts. *CoRR*, abs/2102.12634, 2021. URL https://arxiv.org/abs/2102.12634.

[2] AutoGPTQ. Autogptq: transformers meets autogptq library for lighter and faster quantized inference of llms, 2024. URL https://colab.research.google.com/drive/1_TIrmuKOFhuRRiTWN94iLKUFu6ZX4ceb#scrollTo=H_D9kG_efts3. Accessed: 2025-01-03.

[3] A. Fan, M. Lewis, and Y. Dauphin. Hierarchical neural story generation, 2018. URL https://arxiv.org/abs/1805.04833.

[4] J. Guan, F. Huang, Z. Zhao, X. Zhu, and M. Huang. A Knowledge-Enhanced Pretraining Model for Commonsense Story Generation. *Transactions of the Association for Computational Linguistics*, 8:93–108, Jan. 2020. ISSN 2307-387X. doi: 10.1162/tacl_a_00302. URL https://doi.org/10.1162/tacl_a_00302. _eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00302/1923261/tacl_a_00302.pdf.

[5] Huggingface. Huggingface: Transformers, 2024. URL https://huggingface.co/. Accessed: 2024-11-30.

[6] M. Jumelle. Métriques des llm : mesurer leurs performances, 2024. URL https://blent.ai/blog/a/metriques-llm. Accessed: 2024-11-30.

[7] N. S. Keskar, B. McCann, L. R. Varshney, C. Xiong, and R. Socher. Ctrl: A conditional transformer language model for controllable generation, 2019. URL https://arxiv.org/abs/1909.05858.

[8] Q. Qi, L. Ni, Z. Wang, L. Zhang, J. Liu, and M. Witbrock. Epic-level text generation with llm through auto-prompted reinforcement learning. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2024. doi: 10.1109/IJCNN60899.2024.10650358.

[9] P. Tambwekar, M. Dhuliawala, A. Mehta, L. J. Martin, B. Harrison, and M. O. Riedl. Controllable neural story generation via reinforcement learning. *CoRR*, abs/1809.10736, 2018. URL http://arxiv.org/abs/1809.10736.

[10] Z. Zhao, S. Song, B. Duah, J. Macbeth, S. Carter, M. P. Van, N. S. Bravo, M. Klenk, K. Sick, and A. L. S. Filipowicz. More human than human: Llm-generated narratives outperform human-llm interleaved narratives. In *Proceedings of the 15th Conference on Creativity and Cognition*, CC '23, page 368–370, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701801. doi: 10.1145/3591196.3596612. URL https://doi.org/10.1145/3591196.3596612.