Eesha Pamula
Siddhant Rao
Arnav Devineni
Nishank Gite

Main Goal:
The main goal of the project is to enhance the capability of large language models (LLMs), specifically LLaMA 2, to generate dynamic and personalized stories that are coherent and engaging, based on the user's input, such as character traits, settings, and plot preferences. In our project, we seek to apply fine-tuning techniques to optimize our model towards creating engaging and comprehensive stories.

Main Claims:
The project claims that fine-tuned LLMs can produce structured narratives that adhere to classic short story arcs and interact dynamically with user preference. We also claim that through advanced fine-tuning techniques like quantized Learning Rate Annealing (qLoRA), the tailored narratives can exhibit improved personalization and variability without sacrificing coherence or relevance.

Experiments:
Experiments conducted include the fine-tuning of LLaMA 2 using qLoRA on a specially curated dataset comprising classic short stories and structured narrative datasets. The user preferences are taken into account through meta-data tags introduced during pre-processing, and evaluated during training. Also experimenting with other models, such as Mistral to see if we can appropriately train with the GPU units available, and testing to see if there are any training errors that pop up with different models.

Evaluation Protocol:
We use a combination of the Perplexity scores, and human assessment for evaluation of training data models, and for test/validation data we are using BLEU and ROUGE.

Data:
“This work collects a large dataset of 300K human-written stories paired with writing prompts from an online forum that enables hierarchical story generation. Our dataset enables hierarchical story generation, where the model first generates a premise, and then transforms it into a passage of text.” We are using the same data that is found in the Hierarchical Neural Story Generation paper, but we need to pre-process this data to feed it into our LLM.

Task:
The task is to generate coherent, structured, and personalized short stories that integrate user-provided narrative elements, maintaining the integrity and coherence typical of human-crafted narratives.

Experiments Supporting the Goal/Claims:
The experiments demonstrate the ability of the fine-tuned model to incorporate complex narrative structures and user preferences. By testing different models and based on compute power, determining which one we can use we isolated the appropriate base model. Also experimenting with qLoRA and different quantitative metrics for performance analysis is helping to create a stronger model in terms of cohesive and appealing short story generation.

Limitations Discussed:
The paper discusses the inherent limitations of LLMs, such as potential biases in the generated content and the bounded creativity that relies on the training data. During the experimentation, the demanding resource requirements made it difficult to run some of the large models, leading to system constraints with GPU/CUDA errors.

Strengths of the Paper:
The paper's strength lies in its novel approach to personalized storytelling using LLMs, its rigorous evaluation process combining various quantitative and qualitative measures, and the exploration of fine-tuning techniques tailored to enhancing engaging and understandable narrative generation. Accompanying that, the use of a very large dataset of around 300,000 stories helps ensure proper identification of prompt-story correlations and trends.

Weaknesses of the Paper:
The weaknesses include the challenge of scaling the personalized narrative generation process and the need for further investigation into real-time adaptability and feedback integration. Furthermore, a notable constraint that impacts the paper's comprehensiveness is the limitation in GPU and RAM resources, hindering the exploration of other models such as Mistral. Without these comparisons, we could not understand how the proposed approach fares against alternative methodologies. The initial weakness in the inability to directly use traditional testing loss methods is a weakness, but we are working towards countering that by applying new methods of perplexity, BLEU and ROUGE.

Suggestion for Improvement:
The paper could be improved by extending the diversity of the training datasets, incorporating more real-world user feedback, and conducting additional comparison studies with other models like GPT-2 and Mistral. This would help build a more robust generative model on unique user inputs.

Relevant Related Work:
The project draws on previous research in the areas of generative storytelling, NLP, and user experience design as found in the Hierarchical Neural Story Generation paper. We position ourselves in a similar context by evolving LLM applications for creative content generation by

applying parameter-efficient fine-tuning methods such as qLoRA to a trained LLaMA 2 model, but in a similar vein, people have created story generation LLMs through GPT-2.

Reproducibility:
The project is designed to be reproducible with a detailed methodology with publicly available datasets/frameworks found on HuggingFace and a codebase hosted on Google Collab. By directly running the code on Google Collab you can reproduce our same results.

Clarity of Plots and Paper Text:
All plots and explanations (made thus far) within the paper are clear and interpretable, with well-defined axes and methodological descriptions.

English Quality:
The English used in the paper is correct, clear, and complies with academic standards.

Feedback on TODOs:
The authors need to complete the comparative analysis with LLaMA 2 and further explore the model's interaction with live user feedback. Furthermore, crafting appropriate loss curves and model performance analysis is another TODO.