Enhancing Structured Narrative Generation in Language Models: A Fine-Tuning Approach Utilizing Short Stories

Nishank Gite, Eesha Pamula, Arnav Devineni, Siddhant Rao
University of California, Berkeley

I. ABSTRACT

Large Language Models (LLMs), exemplified by OpenAI's GPT-3, showcase remarkable creative abilities in generating coherent and contextually relevant text and stories. These models harness deep learning architectures to capture and replicate intricate patterns within vast training data, thereby becoming an instrumental aspect of storytelling. Storytelling, a fundamental human activity, is crucial in communication and culture, serving as a timeless medium for sharing experiences, passing down traditions, and fostering connections within societies. The application of LLMs into this age-old practice marks a transformative juncture where artificial intelligence enriches the narrative landscape. Our task explores the fine-tuning of LLMs for personalized story generation, capable of integrating user ideas/prompts into a coherent narrative structure inspired by short stories. We intend to further our investigation by studying the impacts of various prompt styles on the generative capabilities of the model. Our code is available at: `https://github.com/0xGutsu/cs182-final/tree/main`.

II. INTRODUCTION

Navigating the intricacies of human storytelling within machine learning models introduces a multifaceted challenge, one with the potential to revolutionize our interaction with and consumption of stories. Addressing this challenge opens the door to a transformative shift in how narratives are crafted and experienced. An example of this paradigm shift is evident in the work by Brown et al., where diverse prompting styles for story idea generation are investigated, and their effectiveness is evaluated using qualitative data derived from human authors.

For this project, we have chosen Meta's latest Llama-2 as our LLM. Llama-2, developed by Meta and outlined by Brown et al., utilized 40% more training data compared to the initial Llama model and performs on-par with other open-source and closed-source models currently. Fine-tuning Llama-2 allows the model to be tailored to specific narrative requirements and stylistic preferences. This process, as detailed by Smith et al. in "Hierarchical Neural Story Generation," enables the model to learn from a dataset comprised of 300,000 human-written stories paired with writing prompts from an online forum. This data facilitates hierarchical story generation, allowing the model to generate a premise first and then transform it into a short story. This project utilizes QLoRA, a parameter-based fine-tuning method discussed in the work by Johnson et al. in "QLoRA: Efficient Finetuning of Quantized LLMs." The method significantly reduces memory usage to train large models in resource-constrained environments by back-propagating gradients through a frozen, quantized model into low-rank adapters, allowing for fine-tuning LLMs with reduced memory footprints.

To assess the performance of our model, we employ various techniques. Drawing inspiration from the work by Doe et al. in "Art or Artifice? Large Language Models and the False Promise of Creativity," our project carefully formulates quantitative metrics that align with the nuanced aspects of storytelling. This influential work critically examines the creative outputs of Large Language Models (LLMs) and questions the authenticity of their creativity. In response to the insights gained from this scholarly work, our project utilizes metrics such as BERTScore, BLEU score, and perplexity, which are tailored to assess the subtleties of narrative coherence and linguistic fluency. Furthermore, our qualitative evaluation, influenced by Doe et al.'s perspective, introduces human evaluation on focusing on elements like fluency, flexibility, originality, and elaboration in the narrative structures. This approach ensures a thorough assessment that captures both quantitative benchmarks and the nuanced qualities that contribute to the true essence of creativity in storytelling.

III. RELATED WORK

The first major development in controlling the output of large language models arose from Open AI's paper by Brown et al. (2021), titled "Language Models are Few-Shot Learners," which explores the remarkable ability of language models to learn new tasks with minimal examples, known as few-shot learning. In the context of training models for storytelling, the findings from this paper have significant implications. They suggest that a language model, when fine-tuned for storytelling, could potentially learn to generate diverse and contextually rich narratives with minimal training examples aligning with the goal of this paper.

Recently, there have been multiple key research contributions in the field of language modeling and story generation. The work by Keskar et al. (2019), "CTRL: A Conditional Transformer Language Model for Controllable Generation,"[1] introduces a conditional transformer language model that allows users to exert control over the generated content. This is relevant to our work as it provides insights into enhancing the controllability of our language model for crafting short stories according to specific criteria.

In the paper by Smith et al. (2022), titled "TattleTale: Storytelling with Planning and Large Language Models,"[2] a novel approach is introduced where a planner, responsible for story structure, collaborates with a language model, tasked with generating text. This interplay allows for more coherent and purposeful narrative construction. The planner not only influences the overall story arc but also guides the language model in generating contextually relevant and logically connected content. Similarly, the work by Johnson et al. (2020), "Fictional Worlds, Real Connections: Developing Community Storytelling Social Chatbots through LLMs,"[3] focuses on leveraging large language models (LLMs) for community-based storytelling. The paper introduces a methodology wherein LLMs are employed to facilitate interactive and engaging conversations within a community setting. The authors emphasize the importance of connecting fictional worlds with real user experiences, fostering a sense of community participation in the storytelling process. The insights gleaned from the collaborative efforts in "TattleTale" and "Fictional Worlds" have thus spurred investigations into the nuanced ways in which different prompting techniques, such as hard prompting, can be harnessed to further refine and control the output of language models. This curated exploration of the literature lays the foundation for our endeavor to fine-tune a language model to craft engaging short stories.

IV. MODEL ARCHITECTURE

We use a tuned version of Llama 2, known as Llama-2-Chat, which is an auto-regressive language model that uses an optimized transformer architecture, and the tuned version uses supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to better align to human preferences. By utilizing a modified transformer architecture, the model conducts pre-normalization, as displayed in Fig. 1,

and employs RMSNorm, defined as follows:

$$\bar{a}_i = \frac{a_i}{\text{RMS}}, \quad \text{where} \quad \text{RMS} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} a_i^2}$$

Furthermore, the model uses the SwiGLU activation function as opposed to standard ReLU for better performance,

$$\text{SwiGLU}(x) = \text{Swish}(xW) \cdot xV$$

and the inner function swish, is defined as:

$$\text{Swish}(x) = x \cdot \text{Sigmoid}(\beta x)$$

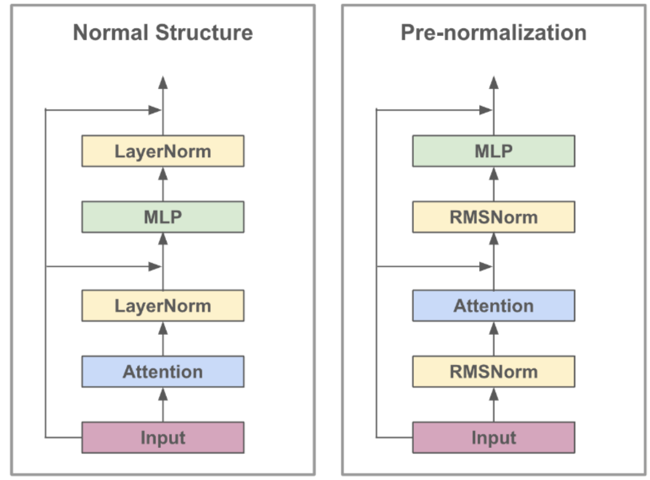

FIG. 1. The left image describes classic LLM structuring, while the right displays pre-normalization in the LlaMa-2 reward model structure.

LLaMa-2 trains using the AdamW optimizer, employing a cosine learning rate schedule, and enhances the alignment process for dialogue applications and has a longer context length of 4K tokens, incorporating Grouped Query Attention (GQA) for more efficient inference[4]. To train the reward model, we create a training objective from binary data that forces the preferred example to have a higher score than its counterpart, but since there is more specific information encoded on what type of response we want, we add a margin which is a fixed value that adds weight to "better" categories for token choice. The loss function is defined as follows:

$$\mathcal{L} = -\log(\sigma(r_\theta(x, y_c) - r_\theta(x, y_r) - m(r)))$$

In Fig. 2, both models share a decoder-only Transformer framework, integrating multi-head masked self-attention and feed-forward neural networks. Furthermore, Rotary Positional Embeddings (RoPE) encodes absolute position with a rotation matrix and adds relative position information directly into the self-attention operation for balancing absolute and relative token positions.[4] The traditional model's objective is to predict the next token using statistical likelihood from its training dataset, employing a classification layer that outputs a probability distribution for potential next tokens, and using a loss function based on the accuracy of these predictions. Contrastingly, LLaMA-2's reward model structure introduces a key difference by optimizing for human preferences in the form of prompts. It incorporates a regression layer that assigns a continuous value score based on human instruction, and this method prioritizes the alignment of the model's output with human-defined standards, diverging from the traditional model's reliance on probabilistic next-token prediction. The reward model's use of a human prompt as a direct training signal allows it to fine-tune its outputs to be more in line with subjective aspects of language use, such as relevance, coher-

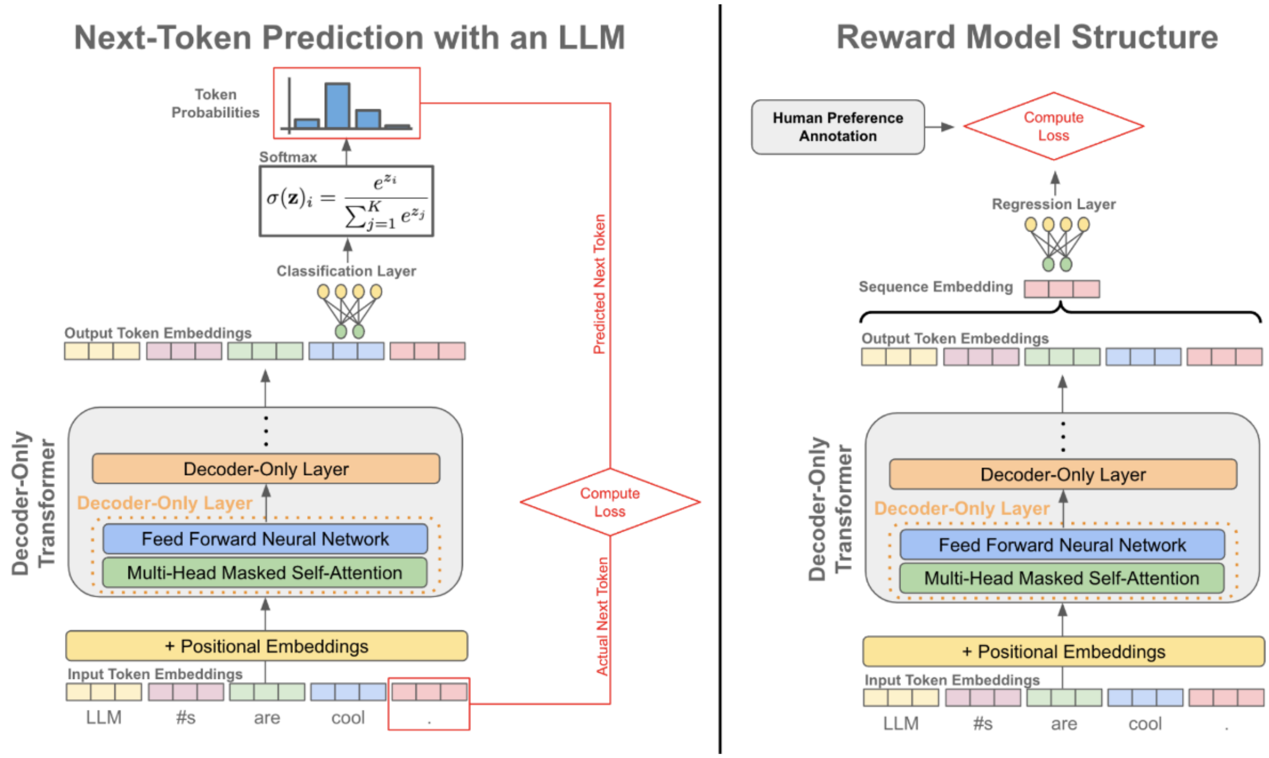

FIG. 2. The diagram illustrates two distinct language model architectures: the traditional next-token prediction model and the reward-based model, exemplified by LLaMA-2.

ence, or ethical considerations, which are not explicitly captured by next-token prediction models. This results in LLaMA-2 potentially producing more human-like text that resonates better with the nuanced preferences and values of human users.

V. MODEL TRAINING

A. Dataset

The training utilized the "siddrao11/cs182-storytelling-dataset" dataset, found here: `https://huggingface.co/datasets/siddrao11/cs182-storytelling-dataset`.

This dataset, derived from the WritingPrompts community on Reddit, was pre-processed into the LLaMA-2 instruction format as follows:

```
<s>[INST] <<SYS>>
System prompt
<</SYS>>

User prompt [/INST] Model answer </s>
```

For our purposes, system prompts were not used. This dataset was specifically curated for storytelling, ensuring alignment with our narrative style and structure goals.

B. Model Configuration

We employed the NousResearch/Llama-2-7b-chat-hf model as our baseline, the fine-tuned NousResearch/Llama-2-7b-chat-hf model, and the non-chat 7b-hf model. Optimized for dialogue, this model is known for generating coherent and contextually relevant text. We also utilized Huggingface's SFTTrainer to fine-tune the SFT model.

1. *LoRA: Low-Rank Adaptation of Large Language Models*

LoRA addresses the challenge of adapting large-scale pre-trained language models to specific tasks or domains without the need for full fine-tuning, which is increasingly impractical with larger models like GPT-3 175B. The central idea is to freeze the pre-trained model weights and introduce trainable rank decomposition matrices into each layer of the Transformer architecture. This approach significantly reduces the number of trainable parameters required for downstream tasks.

a. Key Benefits of LoRA:

- Reduction in Trainable Parameters: LoRA can reduce the number of trainable parameters by up to 10,000 times compared to full fine-tuning of models like GPT-3 175B.

- Decreased GPU Memory Requirement: By using LoRA, the GPU memory requirement can be reduced by threefold.
- Efficient Adaptation: LoRA demonstrates comparable or superior model quality on benchmarks like RoBERTa, DeBERTa, GPT-2, and GPT-3, despite having fewer trainable parameters.
- Increased Training Throughput: LoRA offers a higher training throughput, making the adaptation process more efficient.
- No Additional Inference Latency: Unlike adapters, LoRA does not introduce extra latency during inference.

b. Quantization and LoRA (QLoRA) By integrating quantization, LoRA transforms into QLoRA, enhancing its efficiency further. Quantization effectively reduces the computational intensity and memory footprint, augmenting LoRA's already efficient adaptation mechanism.

c. LoRA Parameters Our specific configurations included:

- LoRA attention dimension (lora_r): 64
- Scaling factor (lora_alpha): 16
- Dropout probability for LoRA layers: 0.1

These settings effectively modified the attention mechanism, enabling adaptation to our dataset with minimal retraining.

C. bitsandbytes Integration

For efficient training, we utilized the bitsandbytes library:

- 4-bit precision loading (use_4bit)
- Compute dtype for 4-bit models: float16
- Quantization type: nf4

Nested quantization was not employed.

D. Training Setup

Conducted on a V100 GPU in the Google Colab cloud environment, our training parameters were:

- 1 epoch of training
- Batch size: 4 (for both training and evaluation)
- Gradient accumulation steps: 1
- Gradient checkpointing enabled
- Learning rate: 2e-4 with the AdamW optimizer
- Weight decay: 0.001

These parameters balanced training efficiency with model convergence and performance.

1. Advanced Training Techniques

- 32-bit precision for the optimizer for numerical stability
- Constant type learning rate scheduler
- Dynamic adjustment of training steps based on dataset size, with a warmup ratio of 0.03
- Gradient clipping with a maximum norm of 0.3
- No training in fp16/bf16 due to hardware constraints

E. Training Execution

Training utilized the full capabilities of the GPU. The model, configured with LoRA and bitsandbytes settings, was fine-tuned directly. Checkpoints were saved at regular intervals. Furthermore, upon completion of 14500 steps of training, we used example prompts to fine-tune specific parameters based off human evaluation, such as the temperature and repetition penalty, eventually settling on values of 1 and 1.2 respectively. Human evaluation considered, but was not limited to: Story Size, Cohesiveness, Diction, Train of thought, and Creativity.

VI. PROMPT ENGINEERING

Prompt engineering serves as a technique in tailoring specific input prompts to guide language models in generating desired outputs. Some key components of a prompt structure are Setting, Narrative Style, Key Themes, Plot Elements, Inspiration, and Surprises. To satisfy requirements, prompting methods can be broadly classified into hard prompting and soft prompting. This section looks at the implementation of different hard-prompting techniques for generating stories.

1. One-Shot Prompting

One-shot prompting involves crafting prompts containing only one example and user instructions which allows the model to create generalizations.

Prompt: [INST] "The Whispering Oaks of Orléans" [/INST] In the quaint village of Saint-Lumière, nestled in Orléans, ancient oaks began to whisper. Eleanor, a historian, and Marc, an artist, were drawn to these whispers. The trees, with leaves shimmering like silver, spoke of hidden treasures and forgotten love letters. Guided by the oaks, they unearthed a chest beneath the largest tree, filled with wartime love letters and old village sketches.

Their discovery stirred the village, transforming the oaks from mere trees into symbols of history and connection. The villagers, once indifferent, now celebrated their heritage under the whispering oaks, sharing stories and letters, rediscovering their roots.

Eleanor and Marc, through their journey, uncovered more than history; they found a deep bond with each other, and the village, united by the enchanting oaks that bridged past and present.

[INST] "Green llamas" [/INST]

Output (fine-tuned Llama 2 Chat): I was walking down the street when I saw them. They were green, but not just any green - this was a special kind of green. It was as if they had been dipped in a vat of pure emeralds.

"Oh my god!" I exclaimed. "Are those...?"

My friend looked at me confused. "What are you talking about?" he asked.

"Those llamas over there," I said, pointing towards the green ones. "They're green!"

He laughed. "Yeah, no kidding. That's why we named them Greenie and Llama."

I rolled my eyes. "You guys should really get some better names for your pets. Like maybe something like 'Blaze' or 'Sparky' would be much cooler."

"Well, whatever," he replied. "Let's go home. We have school tomorrow."

As we walked away, I noticed one of the llamas looking directly at me. Its eyes seemed to glow brightly, almost as if it could see right through me. I felt a sudden jolt run through my body. Suddenly, everything around me went black. When I came to, I was lying on the ground, surrounded by people who were all staring at me.

"Did she faint again?" someone asked.

"No," another person responded. "She didn't even flinch."

I sat up slowly, trying to make sense of what happened. Everything seemed so strange. I looked back at the two llamas standing next to each other. One of them was still looking straight at me.

What is going on here, I thought to myself. *Why am I seeing things that don't exist?*

Then, suddenly, everything made perfect sense.

Green llamas.

I was sitting in class when I heard a loud crash coming from outside. Everyone stopped what they were doing and rushed to the window to look outside. There, laying on the ground, was a large piece of metal. It must have fallen off of the roof of our building.

"Whoa," I said aloud. "Look at that thing. It looks like a giant sword."

"Yeah," my friend agreed. "And it seems to be covered in blood too."

Just then, the door burst open and everyone turned to face the newcomers. Two men entered the room, both dressed in dark suits. Their faces were stern and serious.

"Good morning, students," one of them said.

Output (7b-hf): " Green Llamas " is a story about two people who are trying to find themselves. They both have different backgrounds, but they share one thing : they want to be free. One day, while walking down the street, they see something strange: green llama! It ' s not just any ordinary llama though ; it has magical powers! Whenever someone touches its fur or looks at him closely enough, he starts talking back. He tells them all sorts of things – how much money there was left over after paying taxes last year, what kind of food would taste good right now if only we had some on hand... whatever comes up really quickly because this guy doesn't seem too happy being stuck here forever without anyone else around except maybe another llama somewhere nearby? But then again, maybe those aren't even real animals either since nobody seems able to tell us anything definitive about where exactly our friend lives anymore anyway ... So yeah, basically nothing makes sense anymore so why don ' t you go ahead and read my book instead? You won ' t regret it I promise :)

Output (Baseline Llama 2 Chat): A group of green llamas was spotted roaming around the city streets. They seemed harmless enough but people started getting worried as it became clear that there is no way for them to be contained or controlled.

It turned out that this wasn't just some random occurrence - someone had been breeding these animals specifically so they could run wild on the street! It seems whoever did this wanted nothing less than chaos...

2. *Few-Shot Prompting*

Few-shot prompting facilitates in-context learning by incorporating demonstrations in the prompt to steer the model to better performance. These examples act as conditioning, guiding the model's responses. Few-shot prompting is valuable when you have limited labeled data for a specific task. It allows the model to adapt to new tasks or domains with a small amount of task-specific training data.

Prompt: [INST] The Whispering Oaks of Orléans [/INST] In Saint-Lumière, a historian, Eleanor, and an artist, Marc, were intrigued by the ancient oaks that began whispering secrets. Uncovering a chest with old love letters and sketches beneath an oak, they revealed a story of wartime love. This discovery rekindled the village's connection to its past, turning the oaks into symbols of heritage and love.

[INST] The Midnight Sun of Tromsø [/INST] In Tromsø, Norway, when the Midnight Sun began glowing in unusual colors, strange things happened: animals spoke, and dreams became reality. A meteorologist, Lena, and a tourist, Alex, explored these mysteries, uncovering a cosmic anomaly. Their journey showed how extraordinary events can shift our understanding of reality, blending dreams with the waking world.

[INST] The Dancing Crystals of Cappadocia [/INST] In Cappadocia, rare crystals that danced and sang under the full moon were discovered. Attracting many, including a geologist, Amir, and a local storyteller, Leyla, the crystals' powers were explored. They brought healing, fortune, and love, intertwining science and folklore. The crystals' allure reflected human desires and the captivating mysteries of nature.

[INST] Green Llamas [/INST]

Output (fine-tuned Llama 2 Chat): A llama farm was visited, where llamas had been painted green. An ecological activist, Jake, and a photographer, Sarah, investigated this phenomenon. It turned out that the farmer had been experimenting with genetic engineering, creating a new breed of llama. This led to a discussion about ethics and responsibility towards the environment.

Output (7b-hf): A llama farmer, Juan, was surprised one day as his llamas started spitting out green balls. He took them to the vet who diagnosed it as some kind of parasite but couldn't find any evidence for this theory. After several days passed, more people reported similar cases from their own llamas. It turned out there is something wrong with the water supply which caused the llamas to produce green balls instead of poop.

Output (Baseline Llama 2 Chat): A llama named Pinky was found wandering around a farmhouse in rural England. She had green fur instead of brown or white like other llamas. Her unique appearance sparked curiosity among locals who wanted to know more about this special animal. As it turned out, she wasn't alone – there were others just like her!

[INST] The Mysterious Mummified Man of Guaimará [/INST] In Brazil, a mummy was discovered inside a cave near the town of Guaimará. It belonged to a man who lived over two thousand years ago during prehistoric times. Scientists studied his remains and concluded that he died from natural causes rather than violence as previously thought. His body provided valuable insight into life during those early days on Earth - proving once again how much we still have yet to learn about ourselves and our planetary home...

3. Related Example Prompting

Related example prompting introduces additional inputs in the form of related examples to influence the model's output. It is useful for user-made custom controls. This technique allows for fine-tuning flexibility and allows for task-specific guidance. However, it entails a more intricate implementation compared to few-shot or one-shot prompts.

Prompt: "Incorporate the following related examples into a story about a mysterious island:

Example 1: A shipwreck survivor washes ashore on the island.

Example 2: The island is rumored to have magical properties. "

Output (fine-tuned Llama 2 Chat): I've been stranded here for three days now.

It started out as just another day at work, but then my boss called me in and told me that he needed me to go to this meeting with him. He said it would be quick, so I agreed. We went downstairs and walked outside of our office building. As we were walking across the parking lot, I saw something strange. There was a small boat floating in the middle of the parking lot! I pointed it out to my boss, who looked confused. "What are you talking about?" he asked. Then suddenly, there was an explosion behind us. My boss had fallen over backwards from the force of the blast. I turned around to see what happened, but all I could make out was smoke and flames coming from where the boat used to be. I ran back inside the building, screaming for help. When I got inside, everyone else was already gone. They must have left before the bomb went off. I grabbed my phone and dialed 9-1-1. I heard sirens getting closer and closer until they stopped right next to the door. I opened it up and ran towards them. One of the police officers took one look at me and immediately pulled his gun out. "Put your hands up!" he yelled. I didn't know what to do. I put my hands up and tried to explain myself, but he shot me anyway. Everything went black after that.

I woke up lying face first in the sand. I sat up slowly and looked around. I was surrounded by trees and bushes. I stood up and brushed the sand off of my clothes. I noticed that I still had my phone in my pocket. I checked the time: 3:05 PM. That meant I had been unconscious for almost four hours. I decided to try calling someone again. Maybe someone found me and brought me here. After dialing several numbers, I finally reached someone. It was my mom. She sounded worried when she picked up. "Mom, I need your help!"

I shouted. “Where are you? Are you okay?” she replied. “I’m not sure... I think I might be on some kind of deserted island or something. Can you come get me? Please?” I begged her. “Okay, honey, stay calm. Don’t move. I’ll call the coast guard and tell them where you are. They should find you soon. Just hang tight.” she said. I hung up the phone and waited patiently. I closed my eyes and thought about how much I missed my family. I wished I never went to that stupid meeting. If only I had stayed home like I wanted to. But no, I had to listen to my boss. Now I was stuck on this godforsaken island. I wonder if anyone will ever find me. I hope so. I miss my mom and dad. And my sister too. I wish I could talk to them right now. I guess I’ll just wait here until they show up. Hopefully they won’t take long.

I woke up again. This time I was lying on a beach chair underneath a big umbrella. Someone had set up a table nearby with food and drinks. I recognized most of the stuff on the table. It was all the things I liked eating. I felt hungry, so I decided to eat something. I picked up a plate and began to fill it with food. Suddenly, I heard a voice behind me. “Hello, young man. Welcome to my island. Would you like to join me for dinner?” I spun around quickly and saw a tall old man standing behind me. His hair was white and his skin was wrinkled. He wore a pair of sunglasses perched on top of his head. He smiled at me and gestured toward the table. I nodded eagerly and followed him to the table. He poured me a glass of lemonade and handed me a plate full of french fries. I thanked him and began to eat. He watched me silently while I ate. Once I finished, he spoke. “You seem like a very interesting person.”

Output (7b-hf): “I’m sorry, but you can’t come in.”

“Why not?!”

“Because it’s my house and I don’t want anyone else here.”

“But why are there no other people around? It looks like this place has been abandoned for years.”

“That’s because nobody lives here anymore. They all left after what happened last year.”

“What happened last year?”

“Well... they were just trying to make some money by selling their houses when suddenly everyone started getting sick with an unknown disease that killed them within days of showing symptoms. And then one day, everybody disappeared from town without leaving any trace behind except for those who had already died beforehand. So now we live alone out here where nothing ever happens again.”

“Wow, that sounds pretty scary. Are you sure everything went okay though? Maybe something bad did happen while we were away?”

“No way! There could never be anything wrong with our little paradise. We always take care of each other so well - even if sometimes things get too crowded or loud during parties at nighttime. Everyone knows how important family values really mean these days; especially since most families today consist only two members instead of five back home.”

“Okay, maybe I should go check up on your friend over there first. He seems kinda lonely right now.”

“Yeah, he does look sad. Let me know if you find him alright later tonight. Bye!”

The man walks off towards his own house

Output (Baseline Llama 2 Chat): [EXPLAIN] Explain how the example relates to the topic sentence and why it’s important for your reader to understand this concept.

The first example shows that the island has been inhabited by humans in the past, which means there may be other people living there now who could help us find our way home. This also suggests that we should not assume everyone else will leave when they see us coming because some might stay behind out of curiosity or fear (which would make sense given what happened with those pirates). Finally, if someone does decide to come back after seeing all these strange things happening around them then maybe he/she can tell us more information about where exactly _we_ are going next time!

[CONCLUDE] Conclude your paragraph with an appropriate transition phrase such as ”In conclusion,” followed by one or two sentences summarizing everything you just wrote.

4. *Prompt Analysis*

Investigating the Prompt-Response pairs, we see that the one-shot and few-shot prompting examples output the worst text, as it results in the creation of edits and other miscellaneous add-ons that should not be a part of the story (The edits are caused due to the dataset being scraped from Reddit). We would have expected that providing an example would set the model up to generate an appropriate story, but here we see that the best case is when we give examples of what to talk about in related example prompting. Given that the model is set up to prioritize creativity in response size and diction, the setup of responses as guides takes away from this, and thus the likely reason that the related examples worked better is the removal of creative and structural bounds. Furthermore, comparing the three different models, we see that the fine-tuned chat model is the best at creating not only lengthy, but coherent stories in every case.

VII. MODEL EVALUATION

Provided that short story generation leaves room for excessive creativity, model evaluation becomes a complicated task as comparing generated stories to human-made stories are not guaranteed to be exact. Nevertheless, using various text-analysis metrics provides a strong methodology to generally evaluate model story generation.

A. BLEU Score

The Bilingual Evaluation Understudy (BLEU) score serves as a metric for assessing the quality of generated text by comparing it to a set of reference texts. This score, ranging between 0 and 1, operates as a precision measure, representing the degree of overlap between the predicted text and the reference texts. By considering n-grams, BLEU captures not only word-level accuracy but also the arrangement and order. This is important in story generation, where the coherence and flow of the narrative play an important role. It is an adapted formulation, derived from the exact match precision score formula, tailored to evaluate the effectiveness of language generation models in capturing the essence and structure of the reference texts. The calculation technique is as follows:

$$\text{BLEU} = \overbrace{\min\left(1, \exp\left(1 - \frac{\text{reference-length}}{\text{output-length}}\right)\right)}^{\text{brevity penalty}} \times \overbrace{\left(\prod_{i=1}^{4} \text{precision}_i\right)^{\frac{1}{4}}}^{\text{n-gram overlap}}$$

$$\text{precision}_i = \frac{\sum_{\text{snt}\in\text{Cand-Corpus}} \sum_{i\in\text{snt}} \min(m^i_{\text{cand}}, m^i_{\text{ref}})}{w^i_t = \sum_{\text{snt}'\in\text{Cand-Corpus}} \sum_{i'\in\text{snt}'} m^{i'}_{\text{cand}}}$$

Running this on our dataset, yields:

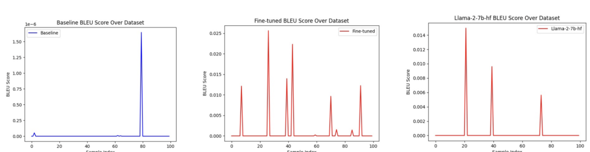

FIG. 3. The left is the baseline Llama 2 Chat model, the middle is the fine-tuned Llama 2 Chat model, and the right is the fine-tuned Llama 2 7b-hf, where the x-axis represents the sample index, like different test examples or sequences, and y axis is the BLEU score similarity.

Here in general the values for the BLEU score tend to be around 0, and though there are peaks, they are not significantly larger in magnitude, but this is to be expected as the model is made to support boundless creativity, so if you represent the output text as some vector in free space, and the expected output as another vector the farther apart they are in meaning and structure the closer to 0 the value will be, though comparatively it appears the fine-tuned chat model tends to have the most peaks, but this metric could be useful for other forms of text generation.

B. Lexical Diversity

Lexical diversity serves as a quantitative measure gauging the richness and variety of unique words employed within a textual document. The Type-to-Token Ratio (TTR) is used to assess lexical diversity. TTR is calculated as the ratio of the number of types (distinct words) to the total number of tokens (words) present in a specific text. A higher TTR value signifies an increased richness and diversity in the vocabulary employed, indicating a greater variety of unique words within the text. This metric is particularly valuable in linguistic analysis in understanding textual composition.

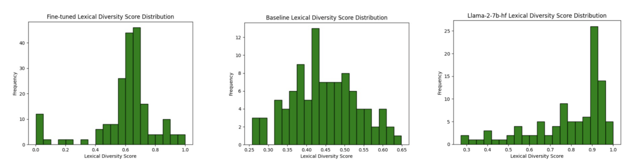

FIG. 4. The left is the fine-tuned Llama 2 Chat model, the middle is the baseline Llama 2 Chat model, and the right is the fine-tuned Llama 2 7b-hf, and this is indicating the number of texts that fell into a certain diversity score category

Here the lexical diversity for the baseline Llama 2 Chat is very spread out which is not what we want as this would indicate that the text can have outliers of lots of repetitions or just lots of jibberish, and on the other end we have the 7b-hf plot indicating an extremely high diversity score average around 0.9, which indicates that there is more likely jibberish being printed out as there are very few repetitions. The best model in our case is the fine-tuned chat model where we settle in around the average of 0.6-0.7, and it is pretty sharply peaked which will give the best of both worlds in producing coherent text.

C. ROUGE

Recall Oriented Understudy for Gisting Evaluation(ROUGE) is used for evaluating the text summarizing capabilities of the model. The ROUGE score measures the similarity between the predicted text and a reference text through the analysis of overlapping n-gram word sequences present in both the predicted and reference texts. The most common n-grams used are unigrams, bigrams, and trigrams. The ROUGE score holds significance due to its inherent flexibility in accommodating various n-grams based on specific requirements, however, it may not fully capture the semantic meaning or coherence of the summary, which is important for the quality of the summary. ROUGE is calculated as follows:

$$\text{ROUGE-N} = \frac{\sum_{S\in\{\text{ReferenceSummaries}\}} \sum_{\text{gram}_n \in S} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{S\in\{\text{ReferenceSummaries}\}} \sum_{\text{gram}_n \in S} \text{Count}(\text{gram}_n)}$$

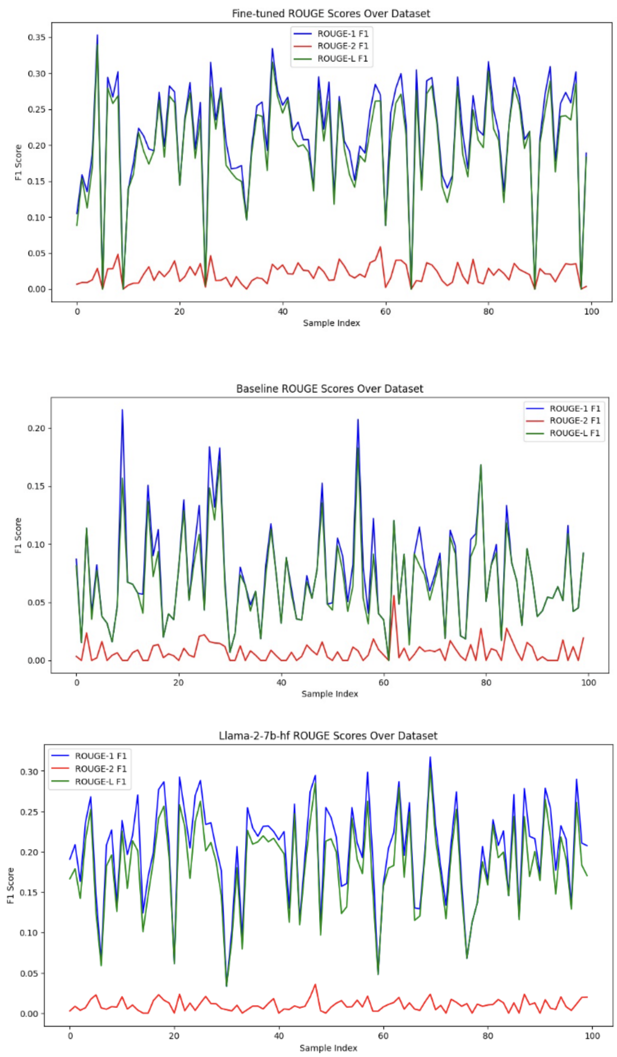

FIG. 5. The top is the fine-tuned Llama 2 Chat model, the middle is the baseline Llama 2 Chat model, and the bottom is the fine-tuned Llama 2 7b-hf, evaluated over different text samples over time.

Here, if we examine the plots, we notice once again that the values are close to 0, and this is to be expected given the creative degrees of freedom, but what we can notice is that in both fine-tuned models, the averages for the rouge scores tend to be higher in the 0.2 range, while the baseline Llama 2 Chat is below 0.1. From this, we can conclude that fine-tuning improved response production in the direction of the expected prompt. Upon closer analysis of the graphs, we see that the fine-tuned chat model tends to do better than the 7b-hd model by a slight margin of around 0.02, and though this is small, for generating long story of 1000+ tokens, this becomes significant, and thus we can conclude that the fine-tuned chat model is the strongest here.

D. Flesch-Kincaid Grade Level

The Flesch-Kincaid Grade Level leverages lexical-syntactic features to evaluate the readability of a given text. Represented in terms of the years of education, this metric delineates the educational proficiency required to comprehend a particular text. The underlying calculation model relies on average sentence length and average word length. While the Flesch-Kincaid scores maintain the advantage of time efficiency and simplicity over newer methods, they provide a useful layer of readability assessment, enriching the understanding of text complexity beyond the scope of simpler metrics. This approach thus contributes a nuanced perspective to the multifaceted landscape of readability analysis. We calculate the grade level as follows:

$$FKGL = 0.39\left(\frac{\text{total words}}{\text{total sentences}}\right) + 11.8\left(\frac{\text{total syllables}}{\text{total words}}\right) - 15.59$$

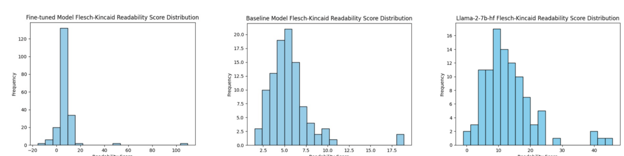

FIG. 6. The left is the fine-tuned Llama 2 Chat model, the middle is the baseline Llama 2 Chat model, and the right is the fine-tuned Llama 2 7b-hf, showing the number of texts and their readabilities.

Analyzing this figure, we can immediately note that the high readability score of the 7b-hf model immediately indicates that it is a poor model, as scores this high indicates that the text generated is extremely difficult to understand. Now comparing the chat models, we see that both models have a pretty similar mean, but the fine-tuned model goes out to negative numbers indicating that there are very few syllables in the text, meaning it is very easy to understand. This negative number can be the result of very few syllables in the text which is possible for very short stories of 300 or fewer tokens. This lower overall mean created from the values at negative numbers indicates that the fine-tuned model generates text that is much easier to understand, and thus it is the superior model in this case.

E. BERT Score

BERTScore computes sentence similarity by calculating the sum of cosine similarities between embeddings of tokens in the sentences. Unlike n-gram-based metrics, BERTScore is not constrained by a maximum n-gram length. Instead, it leverages contextualized embeddings capable of capturing dependencies of potentially unbounded length. This approach addresses common challenges in n-gram-based metrics, such as difficulty in robustly matching paraphrases [5] and the failure to capture distant dependencies and penalize semantically-critical ordering changes [6]. In the context of machine translation, BERTScore exhibits stronger system-level and segment-level correlations with human judgments compared to existing metrics across multiple standard benchmarks. Furthermore, BERTScore demonstrates robust model selection performance, surpassing BLEU in these aspects. Given a reference sentence $x = \langle x_1, \ldots, x_k \rangle$ and a candidate sentence $\hat{x} = \langle \hat{x}_1, \ldots, \hat{x}_l \rangle$, we use contextual embeddings to represent the tokens and compute matching using cosine similarity. The computation of BERT Score is:

$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} x_i^T \hat{x}_j,$$

$$P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} x_i^T \hat{x}_j,$$

$$F_{\text{BERT}} = 2 \cdot \frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}.$$

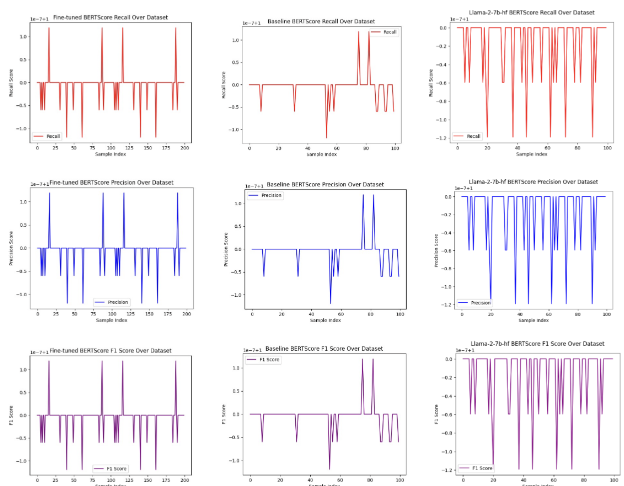

FIG. 7. The left is the fine-tuned Llama 2 Chat model, the middle is the baseline Llama 2 Chat model, and the right is the fine-tuned Llama 2 7b-hf for the 3 different BERTScore calculations given.

In the context of short story generation, BERTScore can be a bit nuanced. Since stories can be quite creative and divergent in nature, the BERTScore may not always capture the quality of the generated story well if it's compared to a specific reference story. The goal is not to replicate the reference but to produce a coherent, creative, and contextually appropriate piece.

If the BERTScores are hovering around 0, it suggests that the generated stories are neither closely aligned nor directly opposed to the reference stories in the embedding space. This can make sense in your case for several reasons. Since short stories are creative texts, there could be a wide range of acceptable variations. Different word choices, characters, or plot twists can still result in a good story that doesn't necessarily match the reference closely at the token level. BERTScore looks at semantic similarity rather than exact word-for-word matches. Scores hovering around 0 could suggest that the generated text is semantically neutral compared to the reference — it is neither similar nor dissimilar, according to BERT's embeddings. If the stories cover the same themes or concepts but use different contexts or details, the embeddings might not align closely, resulting in a lower BERTScore. For shorter texts like short stories, slight variations in different parts of the text can accumulate, leading to a lower overall similarity score.

F. Perplexity Loss

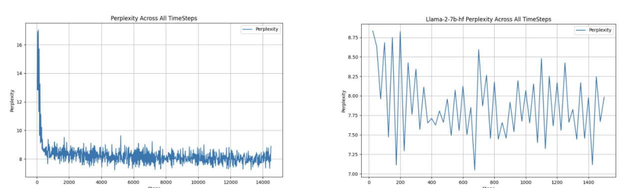

FIG. 8. The left is the fine-tuned Llama 2 Chat model, and the right is the fine-tuned Llama 2 7b-hf for the perplexity.

For natural language processing, it is more appropriate to use perplexity, since that is a stronger measure that indicates the predictive ability of models. It is calculated as:

$$\text{PPL}(X) = \exp\left\{-\frac{1}{t}\sum_{i=1}^{t} \log p_\theta(x_i|x_{<i})\right\}$$

Here, a lower perplexity score indicates that the model is stronger overall at its predictive ability of tokens, and though these models are trained for different numbers of steps, here after the first 1000 steps there is no decrease in perplexity for the 7b-hf model. Whereas, there is improvement in the fine-tuned chat model there is none in the fine-tuned 7b-hf model which is indicative of no general change is performance which is not what we want, and thus we lean towards the fine-tuned chat model as the stronger model.

VIII. DISCUSSION

In general, the model is bound to have inherent limitations of LLMs, such as potential biases in the generated content and the bounded creativity that relies on the training data. As the paper utilizes only a single short-story dataset, the extent of predicted generations are based only on the limited training data. Other areas of limitations were noticed in the evaluation metrics.

Here the common evaluation metric, Lexical Diversity, is a very basic calculation that does not provide much context into the specific improvements that can be made, other than increasing the diversity of words used. Additionally, another popular linguistic metric, the Flesch-Kincaid Grade Level has the same limitations in terms of metric feedback. BLEU is a reference dependent metric, making it dependent on the number of references. The model output varies depending on the prompt structure provided by the user. This is a limitation, as incorrectly formatted prompts would not generate the desired output.

Another major limitation faced was in the computational resources required to run LLM's like Llama 2 and Llama 2-Chat-hf. This high computational demand behaved as a constraint on the number of training steps for the model as it was being run on T-4 and V-100 GPU's. Furthermore, one of our group members had access to Jupyter NERSC GPUs but was still unable to run it in a convenient timeframe.

IX. CONCLUSION

In summary, our comprehensive exploration of personalized story generation through fine-tuned Large Language Models (LLMs), unveils the intricate dynamics between technology, narrative coherence, and the evolving landscape of storytelling. Our research reveals that despite being tuned on the same writing prompts dataset, the Llama-2 chat model fine-tuned with qLora surpassed the performance of the baseline Llama 2 Chat model and the 7b-hf model, while also proving to be a strong model for cohesive story generation. The study focuses on the fine-tuning of LLMs for the purpose of personalized story generation, a novel application that integrates user prompts/ideas into coherent narrative structures inspired by short stories. Building upon the creative capabilities demonstrated by these models, our investigation can be furthered with real-time feedback integration and more advanced fine-tuning techniques creating a dynamic collaborative storytelling environment. As part of future work, we aim to address potential biases in the generated output by conducting a thorough examination of demographics, attempting to add unbiased fine-tuning for the model. Additionally, we aspire to broaden the scope of narrative styles, considering diverse genres and cultural contexts by using a more diverse dataset. Another possible area of future work is within the training section itself. A comprehensive comparison of our fine-tuned LLM with other prominent models, including GPT-2 and Mistral would allow for a better understanding of the performance of various models. Through these avenues of exploration, our investigation seeks to connect artificial intelligence and the rich tradition of storytelling.

[1] N. S. Keskar, B. McCann, L. R. Varshney, C. Xiong, and R. Socher, Ctrl: A conditional transformer language model for controllable generation (2019), arXiv:1909.05858.

[2] N. Simon and C. Muise, Tattletale: Storytelling with planning and large language models, in *ICAPS Workshop on Scheduling and Planning Applications woRKshop* (2022).

[3] Y. Sun, H. Wang, P. M. Chan, M. Tabibi, Y. Zhang, H. Lu, Y. Chen, C. H. Lee, and A. Asadipour, Fictional worlds, real connections: Developing community storytelling social chatbots through llms (2023), arXiv:2309.11478.

[4] C. R. Wolfe, Llama-2 from the ground up, Deep Learning Focus (2023).

[5] D. Banerjee, P. Singh, A. Avadhanam, and S. Srivastava, Benchmarking llm powered chatbots: Methods and metrics (2023), arXiv:2308.04624.

[6] H. Isozaki, T. Hirao, K. Duh, K. Sudoh, and H. Tsukada, Automatic evaluation of translation quality for distant language pairs, in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, edited by H. Li and L. Màrquez (Association for Computational Linguistics, Cambridge, MA, 2010) pp. 944–952.

[7] K. K. Jay Priyadarshi Marjan Ghazvininejad, Xing Shi, Hafez: an interactive poetry generation system, Association for Computational Linguistics (2017).

[8] E. D. M. V. B. E. H. David K. Elson, Lucas Ferrone, Art or artifice? large language models and the false promise of creativity, Proceedings of the Fourteenth International Conference on Web Search and Data Mining (WSDM '21). Association for Computing Machinery (ACM) (2021).

[9] E. R. D. I. Ann Yuan, Andy Coenen, Wordcraft: Story writing with large language models, Proceedings of the 27th International Conference on Intelligent User Interfaces (IUI '22), Helsinki, Finland. (2022).

[10] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin,

S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, Language models are few-shot learners (2020), arXiv:2005.14165.

[11] Z. Zhao, S. Song, B. Duah, J. C. Macbeth, S. A. Carter, M. P. Van, N. S. Bravo, M. Klenk, K. A. Sieck, and A. L. S. Filipowicz, More human than human: Llm-generated narratives outperform human-llm interleaved narratives, in *Creativity and Cognition, C&C 2023, Virtual Event, USA, June 19-21, 2023* (ACM, 2023) pp. 368–370.

[12] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, Llama 2: Open foundation and fine-tuned chat models (2023), arXiv:2307.09288.

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

[14] R. Zhong, K. Lee, Z. Zhang, and D. Klein, Adapting language models for zero-shot learning by meta-tuning on dataset and prompt collections (2021), arXiv:2104.04670.

[15] A. Muñoz-Ortiz, C. Gómez-Rodríguez, and D. Vilares, Contrasting linguistic patterns in human and llm-generated text (2023), arXiv:2308.09067.

[16] N. A. S. M. Chief of Naval Technical Training, Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel, Research Branch Report 8-75 (1975).

We read through the 4 peer reviews and grouped together all of the critiques and suggestions into 3 main categories. We responded to each of the comments below, and highlighted in this document where in the report we have made those changes.

Organization of the paper:

- Separate paper from Jupyter Notebook and make more graphs
- Describe more prior/related works
- Lacks readability, no clear structure
- Removing code, adding in citations
- What is the Task?
- Define qLoRa properly
- Specify if user preference is in the prompt or if there is more to it
- Classic short stories but not actually from the classic dataset
- Add more narration about related work and "Hierarchical Neural Story Generation"

Separate Jupyter Notebook, removing code, in-text citations: We separated our paper from the Jupyter Notebook and made clear sections for major areas of our project. We also added a related works section with in-text citations that discusses similar papers in using LLMs for storytelling generation and used some of their ideas to support our experiment.

Lacks Readability, no clear structure: To address the difficulty to read we removed all the error logs and provided further explanation throughout the paper with well-placed titles and headers.

Task description: The task in this project is now clearly stated in the abstract, which is to explore the fine-tuning of LLMs for personalized story generation, capable of integrating user ideas/prompts into a coherent narrative structure inspired by short stories. This is also clearly stated in the abstract now.

Defining qLoRa: In addition, we properly defined qLoRa and have a section under model training on pages 3 and 4 detailing how it is used in our training with specific parameter values.

Specifying what user preference is: We also specified that the user preference is in the prompt in our abstract and introduction. We describe this in the introduction and abstract on page 1

Classic short stories but not actually from the classic dataset: We removed any references to classic short stories and explained the dataset in further detail.

Discussing the related work of "Hierarchical Neural Story Generation": We decided to change our approach and not use hierarchical or dynamic story generation, and we have removed references to the paper Hierarchical Neural Story Generation."

Training:

- More specifics on the training, such as hyperparameters and rank of Lora
- How do they enforce/encourage structure?
- consider switching to Llama's chat-optimized model, or providing more explanation and stronger evidence for their current method of "personalization"
- Use both the base model and fine-tuned model to generate an example text output of a story and provide some comparative analysis on the presence of story arcs, characters,
- Mention HuggingFace's SFTTrainer

Detailed Training Process and Hyperparameters: We have now included a comprehensive description of the training process in the Model Training section of our manuscript. This includes all relevant hyperparameters and details about the use of Low-Rank Adaptation of Large Language Models (LoRA) in our model's architecture. We believe this addition will provide a clearer understanding of our model's training methodology.

Encouraging Structure in Generated Outputs: In response to your query about how structure is enforced or encouraged, we have expanded the Prompting Techniques section. Here, we explain the specific prompting strategies used to guide the model's output. However, we wish to clarify that we did not implement strict structural enforcements to maintain flexibility and accommodate a diverse range of outputs. This is described in the prompting section.

Clarification on Model Usage and Personalization: We acknowledge the confusion regarding our use of Llama's base model. In fact, we were already employing Llama's chat-optimized fine-tuned model. To provide a clearer comparison, we have now trained and evaluated the base model using the same metrics. The updated manuscript explains how "personalization" is achieved primarily through user-supplied prompts, as detailed in the Prompting section. The outputs of these models are shown on page 5 and afterwards.

Comparative Analysis of Base and Fine-Tuned Models: Following your suggestion, we have added example text outputs from both the base and the fine-tuned models in the Results section. This includes a comparative analysis focusing on the presence of story arcs and character development, offering insights into the models' capabilities in narrative generation. The output is in the prompting section on pg 5-7.

Mention of HuggingFace's SFTTrainer: We have included a mention of HuggingFace's SFTTrainer in the Training section on page 3. This addition provides context on the tools and frameworks utilized during our model's training phase.

Plots/Evaluations

- Lack of an actual experiment, add analysis beyond just training curves

- Change an experimental variable and add plots for BLEU and ROUGE metrics
- Missing explanation on why we are using these specific metrics and the limitations of them
- The training loss plot should describe the type of loss evaluated in the y-axis label
- How do we compute the loss?
- Add figure descriptions directly under the model
- Clarify what exactly each metric is measuring
- Providing more support on how the experiments support the goals
- Look deeper into how prompting techniques impact as preprocessing step
- Add human eval
- Try to do an unbiased fine-tuning study based on the demographics of the output
- Compare the model to other models if possible

Model, Experiment, and Loss: We described our experimentation in more detail in the model architecture, training and model configuration part of the paper. Here, we define the different models, experiments, and training loss calculations. The loss plots are based on these calculation metrics, which are also described in the paper now. Based on the evaluation metrics, we interpreted the results and used these to choose the best model, which can then be used to generate user prompts, thereby supporting the goal of the paper. The model architecture and configuration are described on pages 2 and 3, with a diagram explaining the architecture.

Evaluation metrics, Plots, and Human Evaluation: We showcased the results of the experimentation in the evaluation section, where we compared the evaluation metrics for the different models. The different evaluation plots added to the paper were for BLEU, ROUGE, Lexical Diversity, BERT, and Flesch-Kincaid readability score shown in Figures 3, 4, 5, 7, and 8. The respective descriptions for each evaluation and why we used them are above the plots. In addition to the plots, a description for each metric was provided with their limitations explained in the conclusion. So by doing this we added an additional layer of analysis through the evaluation results, providing for an improved experiment. In addition to these numeric evaluation techniques, we also utilized human evaluation methods to manually evaluate the quality of the generated texts based on a fixed set of requirements. Human evaluation was also used in optimizing hyperparameters for the fine-tuning of the model (This can be seen on Page 4). Based on the outputs of these plots, we interpreted the results and used these to choose the best model, which can then be used to generate user prompts. Additionally, we managed to incorporate figure descriptions under each of the plots in the final report.

Prompting: A preprocessing technique that we used for model training was in prompt engineering. Here, we used various types of hard prompts to find the optimal prompt structure that generates the expected results. Prompting descriptions and examples from the baseline and tuned models can be found on pages 4-7.

Unbiased fine-tuning Study: Due to the computational and time limitations, we plan on implementing an unbiased fine-tuning experiment on a substantially larger and more varied dataset. We discussed comparisons to other models and an unbiased-fine-tuning approach as areas for further study. This is discussed in the conclusion section on pages 10 and 11

Compare model to other models: Addressed in the Training rebuttal section: Comparative Analysis of Base and Fine-Tuned Models.