

# Adapting Large Language Models to Narrative Content

Elio Musacchio<sup>1,\*†</sup>, Lucia Siciliani<sup>2,†</sup>, Pierpaolo Basile<sup>2,†</sup> and Giovanni Semeraro<sup>2</sup>

<sup>1</sup>Italian National PhD Program in Artificial Intelligence, University of Bari Aldo Moro, Bari (ITALY)

<sup>2</sup>Dept. of Computer Science, University of Bari Aldo Moro, Via E. Orabona, 4 - 70125 Bari (ITALY)

## Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities across various domains. However, adapting them to narrative content remains challenging. This paper explores the opportunities in adapting open-source LLMs to narrative contexts, where coherence, plot development, and character consistency are paramount. We investigate existing techniques for adapting and then fine-tuning LLMs on narrative data and propose a solution tailored to the specific demands of narrative generation. Furthermore, we analyze the performance of the proposed approach on the standard dataset WritingPrompts by exploring several corpora for the adaptation step. Moreover, we propose a qualitative evaluation involving human feedback. Results show that the adaptation helps the model improve the generation and accuracy of prompts ranking.

## Keywords

Large Language Models, Narrative Content Generation, Narrative Content Understanding, Generative AI, Artificial Intelligence

## 1. Introduction

In recent years, there have been significant advances in Natural Language Processing (NLP), especially with the rise of Large Language Models (LLMs), which marked a major turning point. LLMs have demonstrated remarkable proficiency in understanding and generating human-like text across several domains and tasks. While these models' initial focus has predominantly been on general language understanding and production, there is now a growing interest in extending their capabilities to more nuanced and complex linguistic tasks. One of the most interesting and challenging domains is creating and understanding creative content.

This paper explores the adaptation of existing open-source LLMs for the purpose of generating and comprehending coherent narratives. Narratives, as structured sequences of events, characters, and interactions, present a unique challenge for language models due to their intricate dependencies on context, temporal order, and thematic coherence. Understanding and effectively generating narrative content requires models to grasp the semantics of individual

---

CREAI 2024 - Workshop on Artificial Intelligence and Creativity, October 19–24, 2024, Santiago de Compostela, Spain

\*Corresponding author.

†These authors contributed equally.

✉ elio.musacchio@uniba.it (E. Musacchio); lucia.siciliani@uniba.it (L. Siciliani); pierpaolo.basile@uniba.it (P. Basile); giovanni.semeraro@uniba.it (G. Semeraro)

🆔 0009-0006-9670-9998 (E. Musacchio); 0000-0002-1438-280X (L. Siciliani); 0000-0002-0545-1105 (P. Basile); 0000-0001-6883-1853 (G. Semeraro)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

sentences, comprehend the broader context and storytelling, and demonstrate commonsense knowledge. Furthermore, the narrative content must also follow the directions given by the user, which poses issues in terms of the controllability of the output [1].

This work aims to investigate the potential of LLMs, with a particular emphasis on open-source models, in adapting their capabilities to narrative content. We aim to explore the challenges associated with narrative adaptation and propose strategies to enhance the performance of these models in generating and understanding stories. We aim to provide open-source models tailored to analyse narrative content to foster further research activities in the field without relying on closed and paid tools.

Our research contributions can be summarized in three points:

- we provide a methodology for adapting an LLM to the narrative content and then fine-tune the adapted model to a specific task;
- we show the applicability of our approach taking into account an open-source LLM (Mistral);
- we provide an extensive evaluation on a standard dataset.

The paper is structured as follows: a detailed description of the methodology is reported in Section 3, while experiments are reported in Section 4. A brief analysis of related work is provided in Section 2 followed by conclusions and future work.

## 2. Related Work

Over the last few years, significant progress has been achieved in the field of Natural Language Generation (NLG) and Natural Language Understanding (NLU) regarding narrative content, witnessing diverse approaches and methodologies. However, despite achieving outstanding results in many NLP tasks, even Deep Learning models perform poorly when addressing the generation and understanding of narrative content like stories and poetry [2].

This result is mainly due to the fact that a story is not simply a concatenation of coherent words or sentences, but it exposes a more complex structure. Creating this structure and keeping it consistent throughout the narration is highly complex. For this reason, researchers have tried to split the story generation task into specific aspects, like event detection [3], and extraction of Characters' networks [4].

Previous research in neural story generation has shown that approaching the story generation task in a hierarchical fashion can improve the structure of the generated content [5]. Later on, several works have focused on incorporating a content planning stage, which proceeds the actual narrative generation, trying to mimic the process adopted by humans to create a story [6, 7, 8].

The complexity of the task increases as the story becomes longer. This is particularly true for stories around a thousand words long, as they approach the length of human-generated short stories found in anthologies. Some works like Yang et al. [9, 10] have specifically focused on long-form text generation.

The aforementioned challenges are inevitably reflected in the evaluation process, making the task even more arduous. First, creating a consistent dataset for evaluating automatically

generated stories is non-trivial, requiring a high cost in terms of time and effort of human experts. Previous works have developed and released datasets for both training and evaluation in the task of story generation. Mostafazadeh et al. [11] released ROC-stories, a carefully crowdsourced dataset consisting of five-sentence long stories focused on realism, coherence and logical succession of events. Fan et al. [5] released WritingPrompts, a dataset consisting of writing prompts and their associated stories written by users, obtained by scraping the WritingPrompts subreddit. Akoury et al. [12] released Storium, a dataset obtained through a collaborative effort with Storium (an online storytelling community), which consists of human-written stories with natural language annotations. However, even when such data is available, establishing effective quantitative metrics is difficult since evaluating stories, whether they are automatically generated or not, involves subjective evaluations. Evaluating stories is not a straightforward process because there are many factors to consider, such as plot, characters, and writing style. All these factors are subjective and can vary depending on the reader’s preferences [2].

A qualitative evaluation procedure involving humans can be performed to overcome this issue, providing them with a survey to complete and collect results from. This allows the retrieval of feedback that considers subjective opinions and other aspects difficult to evaluate using quantitative metrics. Callan and Foster [13] evaluated machine-generated stories on their degree of interest and coherence w.r.t. given writing prompts, while Mori et al. [14] evaluated machine-generated story endings compared to human-written ones and also collected explanations from the users on their choices.

### 3. Methodology

We aim to specialize a generic LLM in generating and understanding narrative content. For this purpose, we suggest a two-step approach. In the first step, the model is adapted to a relatively small dataset of narrative content, and then in the second one, the model is fine-tuned on a specific downstream task in the context of narrative text. In our case, we take into account the task of story generation.

We distinguish between two steps in our train pipeline: *domain adaptation* and *story generation fine-tuning*. For both steps, we collect datasets to perform training.

**Gutenberg Project**<sup>1</sup> is an online library containing over 70,000 ebooks. We collect all the ones in the public domain, using the **Standardized Project Gutenberg Corpus**<sup>2</sup>. This collection process also retrieves the metadata associated with the books (genre, author, and so on). Thanks to the metadata, we are able to filter the dataset using the following criteria:

- only instances in the English language;
- only instances for which the subject field in the metadata contains the “fiction” keyword;
- only instances for which the `authoryearofbirth` field in the metadata is greater than 1800 (the idea is to only keep books that adopt a modern English vocabulary).

---

<sup>1</sup><https://www.gutenberg.org/>

<sup>2</sup>[https://github.com/alex-raw/gutenberg/tree/server\\_fallback](https://github.com/alex-raw/gutenberg/tree/server_fallback)

**BookCorpus** [15] is a dataset consisting of a large corpus of books collected from the *smashwords* website<sup>3</sup>. Despite being widely used in the literature, this dataset has been cited as problematic due to copyright concerns. We opted to exploit the BookCorpus dataset to investigate the second assumption previously presented and test how the presence of more recent works can influence story generation. We also collect metadata for this dataset by using the work proposed in [16] and publicly released on GitHub<sup>4</sup>. Thanks to the metadata, we are able to filter the dataset using the following criteria:

- only instances for which the `Categories` field in the metadata contains the “fiction” keyword;
- only instances for which the `EpubLink` field in the metadata is not empty;
- only instances for which the `Price` field of the metadata is not 0 (avoiding book contents that are not free as reported in [16])

We use both datasets for the adaptation process and split the books’ text to obtain paragraphs. The splitting is done based on the presence of the double newline sequence of characters, and these strings are then considered single instances of data during the domain adaptation step.

Statistics about datasets before and after the filtering procedure are reported in Table 1.

Dataset	Original	Filtered	# of Paragraphs
Gutenberg	63,060	15,280	18,670,952
BookCorpus	413,576	12,999	20,811,688

**Table 1**

Datasets statistics before and after filtering.

For the fine-tuning step on the task of story generation, we use the **WritingPrompts** [5] dataset. This dataset consists of a writing prompt and several possible stories associated with it. WritingPrompts was collected from the homonymous subreddit<sup>5</sup> and released by FAIR (Facebook AI Research), the prompt represents the title of the post, while the stories are the responses the users submitted for the post of that specific prompt. The dataset contains train, test, and validation splits where each instance is represented by a writing prompt (the title of the Reddit post) and a story associated with the former (reply of the Reddit post). We will use the train and test splits for our experiments, which consist respectively of 272,600 and 15,138 instances. However, as it can be seen from Table 2, the quality of the text in the dataset is low (e.g. additional unnecessary white spaces, `<newline>` instead of `\n`). To overcome this issue, we perform the same text pre-processing operations performed by Mao et al. [17], which are:

- **Symbols standardization:** the following rules are used to replace symbols: the `<newline>` symbol is replaced with an actual newline, the double backquote symbol and all other types of quotation marks are replaced by the neutral quote;

<sup>3</sup><https://www.smashwords.com/>

<sup>4</sup><https://github.com/jackbandy/bookcorpus-datasheet>

<sup>5</sup><https://www.reddit.com/r/WritingPrompts>

<p><b>Writing Prompt:</b> [ WP ] You 've finally managed to discover the secret to immortality . Suddenly , Death appears before you , hands you a business card , and says , “ When you realize living forever sucks , call this number , I 've got a job offer for you . ”</p>
<p>So many times have I walked on ruins , the remainings of places that I loved and got used to.. At first I was scared , each time I could feel my city , my current generation collapse , break into the black hole that thrives within it , I could feel humanity , the way I 'm able to feel my body.. After a few hundred years , the pattern became obvious , no longer the war and damage that would devastate me over and over again in the far past was effecting me so dominantly . &lt;newline&gt; It 's funny , but I felt as if after gaining what I desired so long , what I have lived for my entire life , only then , when I achieved immortality I started truly aging .</p> <p>....</p>

**Table 2**  
Example of a raw story from the Writing Prompts dataset

- **Removal of redundant white spaces:** all white spaces before a punctuation mark and a word are removed, as well as instances when more than one white space occurs;
- **Removal of WritingPrompts tags:** in the WritingPrompts subreddit, each post is identified by a tag (e.g. WP = writing prompt, OT = off-topic, ...). In the original scraping process, these tags were kept in the prompt text, therefore they are removed through a simple regex matching operation.

### 3.1. Implementation Details

We decided to use **Mistral-7b** [18] as our base LLM since it has shown remarkable capabilities, being able to surpass the 13b version of LLaMa 2 [19] on all the tested benchmarks.

We consider several configurations according to our assumptions:

- Fine-tune Mistral-7b on WritingPrompts only;
- Adapt Mistral-7b on Gutenberg and then fine-tune on WritingPrompts;
- Adapt Mistral-7b on BookCorpus and then fine-tune on WritingPrompts;
- Adapt Mistral-7b on both Gutenberg and BookCorpus, and then fine-tune on WritingPrompts.

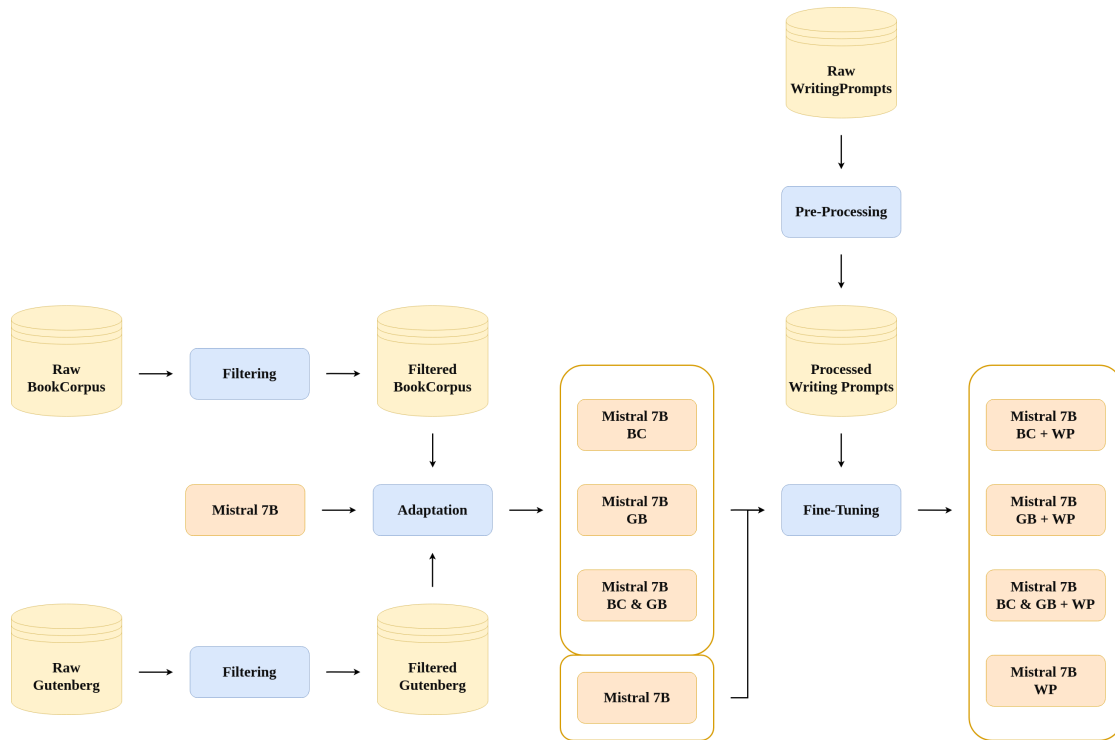
A diagram representing the entire pipeline with the previously described pre-processing steps and these training configurations is presented in Figure 1.

For all configurations, we used the official checkpoint released on HuggingFace by MistralAI<sup>6</sup>, we load it using *flash attention 2* [20] for the attention mechanism implementation. The training procedure was carried out on 4 A100 64 GB GPUs, using DeepSpeed Zero stage 3<sup>7</sup> for parallelism. As for the training arguments, we follow the ones that have been used by HuggingFace<sup>8</sup>, using LoRA [21] to further optimize the trade-off between performance and train time. The only difference between the configurations is the number of training steps. For the fine-tuning process, we train for one entire epoch (covering all instances of the dataset); in the adaptation

<sup>6</sup><https://huggingface.co/mistralai/Mistral-7B-v0.1/commit/26bca36bde8333b5d7f72e9ed20ccda6a618af24>

<sup>7</sup><https://www.deepspeed.ai/>

<sup>8</sup>[https://github.com/pacman100/DHS-LLM-Workshop/blob/4e41ee0a3228d0a34c812f066b1ae7737fa8ae9f/chat\\_assistant/training/run\\_peft\\_deepspeed.sh](https://github.com/pacman100/DHS-LLM-Workshop/blob/4e41ee0a3228d0a34c812f066b1ae7737fa8ae9f/chat_assistant/training/run_peft_deepspeed.sh)



**Figure 1:** Diagram representing the pipeline that was applied

process, we train for 2,500 steps in total. This is due to the substantial difference in the dataset size between adaptation and fine-tuning.

During fine-tuning and inference on WritingPrompts, we use the following prompt inspired by the one used for ALPACA [22], where text in-between curly brackets are variables that are replaced by the corresponding text at train and inference time:

Listing 1: Instruction template

```

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

### Instruction:
Write a story for the writing prompt provided as input

### Input:
{writing_prompt}

### Response:
{story}

```

At train time, both the writing prompt and the story are given as input in the prompt presented in Listing 1, while at inference time, the story is removed and generated by the model. Therefore, the goal is to generate a story influenced by the context provided as input in the prompt.

**Writing Prompt:** A story about a dancer who tries to win the National championship.

This was his chance.

Edgie could see it all happening as he stepped out on stage. The blur of the crowd. Heads bobbing. Camera lenses glistening. \*Boom\* \*Boom\* \*Boom\*

His powerful bass drum thumped out the opening lines.

The crowd grew silent as Edgie's virtuoso began to flourish. The spins. The leaps. The incredible acrobatic feats, Edgie wasn't sure how his legs weren't tired. But they weren't. He was perfect.

Then Edgie's comrade jumped on stage. Edgie's heart sank. Jack was his biggest rival. Together, the two had been slated to go to the national championships.

Jack wasn't dancing tonight though. He had that same look he'd had back in elementary school when he learned that candy had less calories than fruit. He'd glare at anyone who even whispered the word "banana" and never even eat it. He was \*taking\* the trophy this year.

Edgie hadn't realized that he had slumped over on stage. Looking back up at Jack, he saw he had gone out of sync.

\*Jack's dad\*, Edgie thought.

Then he blacked out.

He woke up in an infirmary. He couldn't move. Doctors were rushing around him and muttering his condition to other doctors. He couldn't hear them over the echoing drum in his mind.

Why had it all gone so wrong?

"Edgie. My name is Doctor Baker. I'm going to be able to give you back what's important. By the end of this treatment, your heart rate will be back down to normal, your blood pressure will be lower, and your pulse will be reduced. But I'm not going to be able to give you your dancing back.

"What do you mean?"

"Your pulse was almost 200 beats per minute. Your heart has had to grow extra muscle to keep up with the workload. It just can't function at the rate it was beating last night. We're going to have to reduce it, but that means you won't be able to get that heart-pounding, euphoric, ecstatic rush that you get every time you dance. That's something that's permanent."

As Edgie looked over at his mother, she was talking to the doctor.

"But his rhythm was so good. I was so proud."

"You were doing great, right up until Jack came on. His dad, Tim, wasn't that bad at dancing either. Edgie could never beat Jack though. Every time they went head to head, it was over. And that's what it was going to be like this year too. Jack's getting the trophy.

Edgie had one more chance to win. He knew it wouldn't be enough. He knew he'd get the title, but there would be one moment that would hold the same importance as the final step he took towards victory.

He knew this because Jack had just been wheeled into the room with Edgie. He was even more slouched over than he had been before. He kept his head down the entire time and refused to meet Edgie's gaze. Jack had already lost.

**Table 3**

Example of story generated by our WP model

We release both code<sup>9</sup> and models<sup>10</sup> to allow reproducibility of our experiments.

## 4. Evaluation

We evaluate our generated stories using both quantitative and qualitative measures.

### 4.1. Quantitative Evaluation

For the quantitative evaluation metrics, we generate stories using the prompt structure and the data from the WritingPrompts test split. Following the evaluation criteria presented in [5], we do not use comparison metrics like BLEU or ROUGE since the aim of the generation task should be to write novel and viable stories.

To compare our results with another similar work, we consider what has been done in [17] using GPT-2 [23] for story generation. The authors perform a domain adaptation step

<sup>9</sup>[https://github.com/m-elio/Narrative\\_Content\\_LLMs](https://github.com/m-elio/Narrative_Content_LLMs)

<sup>10</sup><https://huggingface.co/collections/m-elio/story-generation-models-66cf23357d8a3552a3f11785>



<p><b>Writing Prompt:</b> A story about a dancer who tries to win the National championship.</p> <p>This event happens a bit regularly, and one of the main winners is the one and only Riley. Every year it seems, a lot of people go out and learn to dance. Riley decided to try a few years back, but only to come in first place. The crowd went wild, and not much after that, she decided to go into the college of the yearly winner. Only two colleges are available, and each one competes in the National event that is presented at a final, annual tournament. Riley, now twenty-four, has been winning for so long, that she has not much to go up against. Riley has been living with her grandparents for the majority of her life, as her parents died in a car crash ten years ago. Her guardians, now in their early 70's, took it upon themselves to raise her properly. After being out of college for five years, Riley had chosen to go to the college of Miranda. Miranda is also the college of Jennifer, a slightly younger, but experienced dancer. The two dancers had been in competition for the past five years, and everytime they face off, it has been a competition to see who would reign victorious in the end. This year, after the fifteen-rounds of dancing, it was clear which college would be brought home a new trophy. Miranda. "No, no no no no, what. How did we lose. What about Riley, she's supposed to win," screams the President of the College, Vance. "Hey, take it easy. Lets find the winner, maybe there's an error. " Oh my God, I can't believe it. We lost... Riley." "It doesn't matter, who was the best dancer?" "Oh well the victorious one is Jennifer. We should congratulate her." " Congratulations Jennifer, you will now be representing Miranda in the Nationals!"</p>
--

**Table 4**  
Example of story generated by our GB&BC + WP model

(on BookCorpus) and fine-tuning strategy (on various story datasets) using GPT-2. Since the authors worked with a maximum sequence length of 1024 for both the train and evaluation phase (which is the limit for the model), we also evaluate our models with a maximum sequence length of 1024 to keep the results comparable. However, since Mistral-7b has a maximum sequence length of 4096, we also try evaluating 2048 as the maximum length. We expect the latter to perform better, considering it is also the truncation length value we used during the training phase.

The metrics that we use are the following:

- **Perplexity**<sup>11</sup> can be defined as the exponential of the cross-entropy between the model predictions and the actual data. Given that this metric is influenced by the tokenization strategy and our model predicts tokens and not words, we adapt to Mistral-7b's tokenizer the code released by Mao et al. [17] to compute word-level perplexity for GPT-2;
- **Prompt Ranking Accuracy** is a metric used to measure the dependence of a prompt on an output story. Specifically, given 1 prompt and 1 story associated with it, 9 random prompts are sampled from the test split. Then, the percentage of cases where the true prompt is most likely to generate the story is measured (that is, when the average cross-entropy loss of prompt+story is minimum for the actual prompt w.r.t. other prompts). In our experiments, we use 1,000 randomly sampled correct prompts to calculate the metric, following what was also done by Mao et al. [17]. Since each prompt can be associated with more than one story, we ensure that all 10 prompts (the correct and the others) are different.

Table 5 reports the evaluation results. Domain adaptation using the Gutenberg or BookCorpus corpora did not improve results with respect to fine-tuning on WritingPrompts only in terms of Perplexity, and only a small boost was obtained for Prompt Ranking Accuracy. We assume

<sup>11</sup><https://huggingface.co/docs/transformers/perplexity>



Model	Max Length 1024			Max Length 2048		
	SW PPL ↓	W PPL ↓	PRA ↑	SW PPL ↓	W PPL ↓	PRA ↑
Best Results in [17]	20.78	29.52	80.6%	X	X	X
MISTRAL-7B	11.82	27.07	80.5%	11.62	26.64	86.3%
MISTRAL-7B + WP	<b>8.11</b>	<b>24.19</b>	87.3%	<b>8.09</b>	<b>23.87</b>	94%
MISTRAL-7B + GB + WP	8.18	24.23	87.3%	8.18	23.93	93.3%
MISTRAL-7B + BC + WP	8.13	24.32	87.1%	8.12	24.01	<b>94.2%</b>
MISTRAL-7B + GB&BC + WP	8.16	24.24	<b>87.4%</b>	8.16	23.93	<b>94.2%</b>

**Table 5**

Quantitative Results Table. The best result for the associated max length is reported in bold for each metric. Models legend: **WP** refers to the WritingPrompts dataset, **GB** to the Gutenberg dataset, **BC** to the BookCorpus dataset and **GB&BC** to the dataset obtained by combining Gutenberg and BookCorpus. Metrics legend: **SW PPL** refers to Sub-Word Perplexity, **W PPL** refers to Word Perplexity, **PRA** refers to Prompt Ranking Accuracy. The upper and lower pointing arrows are used to indicate whether the higher or lower score is better.

that this is due to the nature of public domain books, which tend to be older works, while WritingPrompts is a dataset obtained by Reddit, where the language used is modern.

As expected, we observe better results when a greater max length (2048) is used during the test. In particular, we significantly improved the Prompt Ranking Accuracy (PRA).

However, while these metrics are useful for understanding what the models have learned, they fail to capture the richness and originality of the stories that the models can generate. Hence, we decided to test our models with humans in a qualitative evaluation procedure.

## 4.2. Qualitative Evaluation

For the qualitative evaluation procedure, we setup a website using a GRADIO<sup>12</sup> interface to allow users to interact with our models. Figure 2 shows an example of interaction with the interface. Users can submit writing prompts to the website, and the model generates a story based on that input. After the story is generated, users are asked to answer three questions in *Likert scale* from 1 to 5 to evaluate the quality of the generated story. The questions ask the user to evaluate the story in terms of three different aspects:

- **Readability:** grammatical correctness and fluency of the text of the generated story;
- **Coherence:** coherence of the generated story with respect to the writing prompt provided as input;
- **Creativity:** the degree of novelty and uniqueness for the characters and the plot of the generated story.

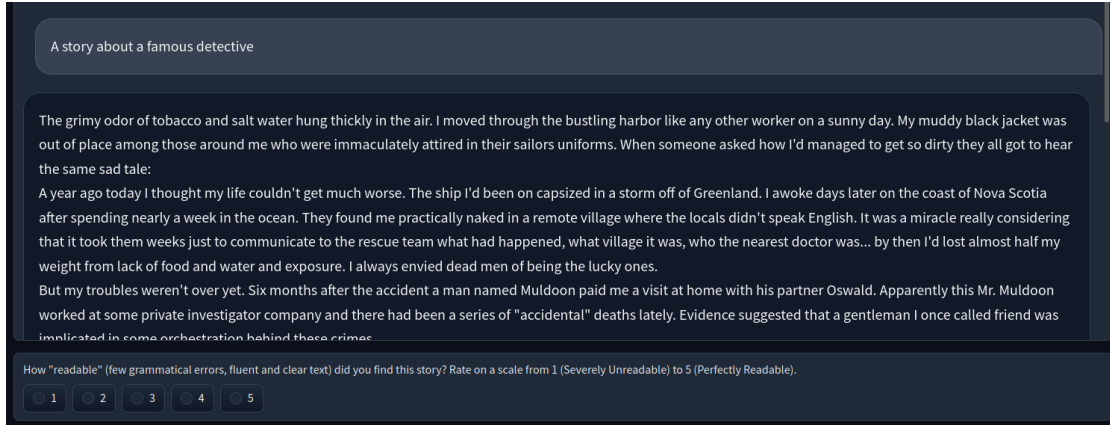
For all metrics, a score value of 1 represents the lowest possible score, while 5 is the highest. We define these metrics to check the models’ capability to generate grammatically correct text that respects the writing prompt given as input, which is new w.r.t. classic stories. We use the MISTRAL-7B + WP and the MISTRAL-7B + GB&BC + WP with 2048 as the max length since they

<sup>12</sup><https://www.gradio.app/>

Model	Readability	Coherence	Creativity
MISTRAL-7B + WP	3.40	2.53	3.27
MISTRAL-7B + GB&BC + WP	3.27	2.97	3.23

**Table 6**

Qualitative Results Table. The average of the scores obtained for each generated story is reported for each metric.



**Figure 2:** Gradio Interface generation example

obtained the best results for the quantitative evaluation procedure. The goal of the qualitative evaluation is to confirm the hypotheses obtained in the quantitative evaluation phase and fully understand whether there is a significant difference when using a domain-adapted model.

We set up the platform so that only one model is accessible simultaneously. To collect feedback, we share a link to access the interface with people who willingly agreed to participate in the study. The participants are instructed to freely access the site when they prefer and to send writing prompts of their own creation. After enough time had passed, we switched the model being used on the platform to the second one. Due to the nature of the experimental settings, the people and the submitted prompts differ between the two tested models. In total, 30 requests are obtained for the two models, we report the results of this evaluation phase in Table 6, where the final score is computed as the average of the 30 scores in range 1 to 5 for each question.

We perform the **Mann-Whitney U** statistical test on the obtained results for each metric. We find a p-value greater than 0.05 for all metrics, meaning that the average differences are not statistically significant. We believe this is due to the linguistic difference in the adaptation and instruction-tuning datasets. As previously mentioned, public domain books have a different writing structure than writing samples written by Reddit users. This result suggests that stories generated by the model that has been domain-adapted still follow the linguistic style and structure that has been learned during the instruction-tuning phase.

## 5. Conclusions and Future Work

We provide a methodology for adapting a generic LLM to the narrative domains using a rich text collection. Then, the adapted model is fine-tuned to perform the story generation task w.r.t. writing prompt provided as input. We consider the story generation task by exploiting the WritingPrompts dataset, which consists of Reddit posts with a writing prompt and the stories sent as responses to such posts.

Quantitative results show that the adaptation can improve Prompt Ranking Accuracy if a dataset of books is used during the adaptation step, hinting that the model is more capable of respecting the writing prompt provided as input. Moreover, we observe an improvement in performance when a greater max length supported by the model is used during the test. In particular, we observe a remarkable gain in prompt ranking accuracy.

We also perform a qualitative evaluation involving human feedback, using two of the best performing models (according to the quantitative evaluation step). After performing statistical tests of the Likert scale results, we find that the average differences for all metrics are not statistically significant between the two models.

In this work, we didn't consider other methods for narrative content generation or other tasks (e.g. generation of a narrative character description from a summary of traits). Furthermore, no other training strategies or hyperparameters (e.g. LoRA rank) were tested. Hence, in future work, we plan to test other open-source LLMs, investigate further datasets and techniques for both the adaptation and fine-tuning steps, and evaluate other tasks related to the narrative content.

## Ethics Statement

It is important to note that while we tried to use as many public domain works as possible, some of the data processed in these experiments (specifically, the BookCorpus dataset) may have been subject to copyright restrictions.

Furthermore, the WritingPrompts dataset is a collection of Reddit posts obtained through web scraping. While the posts are public and accessible, users were never asked for consent to publish a dataset containing messages written by them. In Krotov et al. [24], the authors underline how datasets obtained through web scraping techniques may result in unintended harm to others. We also underline that there is no original author information in our experiments, and there won't be any evidence of the source of the generated text.

This work used these resources for research purposes only. Obtained materials should be used for research purposes only.

## Acknowledgments

We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), Spoke 6 - Symbiotic AI (CUP H97G22000210007) under the NRRP MUR program funded by the NextGenerationEU.

## References

- [1] A. Alabdulkarim, S. Li, X. Peng, Automatic story generation: Challenges and attempts, arXiv preprint arXiv:2102.12634 (2021).
- [2] A. I. Alhussain, A. M. Azmi, Automatic story generation: a survey of approaches, *ACM Computing Surveys (CSUR)* 54 (2021) 1–38.
- [3] Y. Chen, L. Xu, K. Liu, D. Zeng, J. Zhao, Event extraction via dynamic multi-pooling convolutional neural networks, in: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015, pp. 167–176.
- [4] V. Labatut, X. Bost, Extraction and analysis of fictional character networks: A survey, *ACM Computing Surveys (CSUR)* 52 (2019) 1–40.
- [5] A. Fan, M. Lewis, Y. Dauphin, Hierarchical neural story generation, arXiv preprint arXiv:1805.04833 (2018).
- [6] L. Martin, P. Ammanabrolu, X. Wang, W. Hancock, S. Singh, B. Harrison, M. Riedl, Event representations for automated story generation with deep neural nets, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- [7] A. Fan, M. Lewis, Y. Dauphin, Strategies for structuring story generation, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2650–2660.
- [8] J. Xu, X. Ren, Y. Zhang, Q. Zeng, X. Cai, X. Sun, A skeleton-based model for promoting coherence among sentences in narrative story generation, arXiv preprint arXiv:1808.06945 (2018).
- [9] K. Yang, Y. Tian, N. Peng, D. Klein, Re3: Generating longer stories with recursive reprompting and revision, arXiv preprint arXiv:2210.06774 (2022).
- [10] K. Yang, D. Klein, N. Peng, Y. Tian, Doc: Improving long story coherence with detailed outline control, arXiv preprint arXiv:2212.10077 (2022).
- [11] N. Mostafazadeh, N. Chambers, X. He, D. Parikh, D. Batra, L. Vanderwende, P. Kohli, J. Allen, A corpus and evaluation framework for deeper understanding of commonsense stories, arXiv preprint arXiv:1604.01696 (2016).
- [12] N. Akoury, S. Wang, J. Whiting, S. Hood, N. Peng, M. Iyyer, Storium: A dataset and evaluation platform for machine-in-the-loop story generation, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 6470–6484.
- [13] D. Callan, J. Foster, Evaluation of interest and coherence in machine generated stories., in: *AICS*, 2021, pp. 212–223.
- [14] Y. Mori, H. Yamane, Y. Mukuta, T. Harada, Toward a better story end: Collecting human evaluation with reasons, in: *Proceedings of the 12th International Conference on Natural Language Generation*, 2019, pp. 383–390.
- [15] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, S. Fidler, Aligning books and movies: Towards story-like visual explanations by watching movies and reading books, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 19–27.
- [16] J. Bandy, N. Vincent, Addressing ”documentation debt” in machine learning research:

A retrospective datasheet for bookcorpus, arXiv preprint arXiv:2105.05241 (2021). URL: <https://arxiv.org/abs/2105.05241>.

- [17] H. H. Mao, B. P. Majumder, J. McAuley, G. Cottrell, Improving neural story generation by targeted common sense grounding, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 5988–5993. URL: <https://aclanthology.org/D19-1615>. doi:10.18653/v1/D19-1615.
- [18] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7b, arXiv preprint arXiv:2310.06825 (2023).
- [19] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).
- [20] T. Dao, FlashAttention-2: Faster attention with better parallelism and work partitioning, arXiv preprint arXiv:2307.08691 (2023).
- [21] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, arXiv preprint arXiv:2106.09685 (2021).
- [22] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Stanford alpaca: An instruction-following llama model, [https://github.com/tatsu-lab/stanford\\_alpaca](https://github.com/tatsu-lab/stanford_alpaca), 2023.
- [23] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.
- [24] V. Krotov, L. Johnson, L. Silva, Tutorial: Legality and ethics of web scraping, Communications of the Association for Information Systems 47 (2020) 22.