A Report on LSG 2024: LLM Fine-Tuning for Fictional Stories Generation

Daria Seredina
National Research University Higher School of Economics, Saint-Petersburg, 194100
daseredina@edu.hse.ru

Abstract

Our methodology centers around fine-tuning a large language model (LLM), leveraging supervised learning to produce fictional text. Our model was trained on a dataset crafted from a collection of public domain books sourced from Project Gutenberg, which underwent thorough processing. The final fictional text was generated in response to a set of prompts provided in the baseline. Our approach was evaluated using a combination of automatic and human assessments, ensuring a comprehensive evaluation of our model's performance.

1 Introduction

The increasing capabilities of machine learning have paved the way for generating various types of content using LLMs. Prompt-engineering methods, such as those proposed by Sanh et al. (2021), have demonstrated potential in creating fictional texts, but still require human oversight to produce coherent and engaging narratives (Guan et al., 2022). To overcome this limitation, when participating in the shared task of human-like long story generation, LSG Challenge (Mikhaylovskiy, 2023) we decided to explore a hybrid approach, integrating fine-tuning and prompt-engineering techniques to enhance long story generation results.

We chose to fine-tune the Mistral-7B-Instruct-v0.2-GPTQ (Jiang et al., 2023) LLM, aiming to make plausible fictional text generation possible. We define the following traits of the plausible fictional text:

- Metaphors and expressive means, complicated vocabulary, emotional depth, stylistic sophistication.
- Compliance with literary and stylistic norms established in various literary genres and directions.
- Convincing characters, peculiar storylines and conflicts that can interest and capture the reader's attention.
- Individuality and originality of the text, which allows the reader to learn new facts, feel emotions and get a unique reading experience.

In this context, fine-tuning of the model allows us to create a more advanced and adapted system for generating a literary text that takes into account the peculiarities of the literary fantasy genre.

2 Pipeline

To participate in the shared task, we leveraged a fine-tuned model and a custom pipeline to generate a comprehensive Harry Potter fanfiction. Our pipeline employs a hierarchical approach to narrative generation, utilizing prompting to create a lengthy and engaging story. The pipeline consists of three key stages:

- We initiate the process by loading the provided prompt and using an untrained Mistral-7B-Instruct-v0.2-GPTQ model to generate a high-level outline of the narrative similarly to the LSG Challenge baseline (Migal et al., 2024) and previous work (Lee et al., 2024, Sun et al., 2022). This outline serves as a roadmap, defining

Proceedings of the 17th International Natural Language Generation Conference: Generation Challenges, pages 123–127
September 23 –27, 2024.

the most critical events that will unfold in the story, including the setup, climax, resolution, and conclusion.

- With the narrative skeleton in place, we generate a more detailed content outline, comprising chapters that correspond to each part of the story. Each chapter is accompanied by a concise plot description, outlining the main events that will happen.

- Once we have a satisfactory chapter outline, we utilize our fine-tuned model (Mistral-7B-Instruct-v0.2-GPTQ) to generate the actual narrative, bringing the story to life.

3 Fine-tuning Approach

For fine-tuning the model, we employed a hybrid approach that combines elements of Supervised Learning (SL) and Self-Supervised Learning (SSL). Specifically, our dataset consisted of pairs of input data (brief summaries of a book chunks) and their associated labels (original book excerpts), which the model used to adjust its parameters and improve its predictive accuracy. This supervised learning aspect allowed the model to learn the mapping between input data and target outputs. However, the task of generating text based on brief summaries does not provide explicit labels for every possible output. Instead, the model must learn to generate text by leveraging the internal relationships between the input data and the desired output, which is a characteristic of self-supervised learning. We chose this hybrid approach due to its effectiveness in optimizing model performance when a clear relationship between inputs and desired outputs can be established.

To achieve this, we constructed a dataset that conforms to the following structure:

- "Outline": a concise summary of the narrative that the model should utilize to generate a fictional text.

- "Reference text": an exemplary text that serves as a benchmark for a well-crafted fictional text.

- "Instruction": a specific prompt that guides the model on how to integrate the " outline" and "reference text" to produce a coherent output.

The "outline" was derived from the original text chunk and distilled into a condensed summary. Consequently, the primary objective for the model during fine-tuning was to reconstruct the original text chunk with fidelity.

4 Dataset

The dataset was constructed using a part of the vast digital collection of public domain books provided by Project Gutenberg (Project Gutenberg, 2016). This collection comprises a broad spectrum of classic literature. To create the dataset, we selected over 500 books from the Project Gutenberg collection, starting from the first available ID, while striving to retain only books containing fictional works and discarding non-fiction texts, as

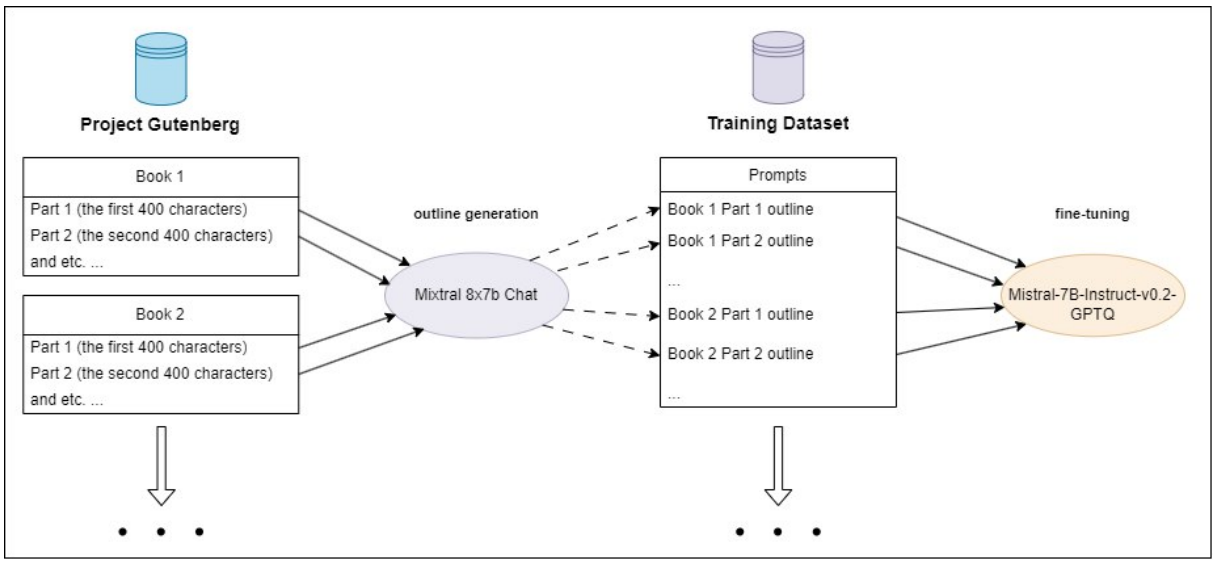

Figure 1: Dataset generation

our primary objective was to fine-tune a model capable of generating artistic text. These books were converted into text format and processed, enabling us to extract plotlines for generating concise summaries or "outlines" with Mixtral 8x7B model (Jiang et al., 2023). To prepare the data for model training, we transformed it into the required format by adding special tokens, the specific prompt, and the reference text, as follows:

"[INST] You are a famous writer and you are writing a book. Now you are going to write one chapter of your book according to this plot: In the beginning, <…>. Add as many descriptions, dialogues, feelings of the main characters and other things as possible. [/INST] Try to write your text as well as this one: *Reference text*."

The final version of the dataset was used for training the model. The dataset creation process is shown in Figure 1.

5 Parameters

For fine-tuning, we utilized a LoraConfig, which includes the following parameters:

- r = 8 - This is a hyperparameter value that determines the vector size for each layer of the model.
- lora_alpha = 32 - Defines the learning scale for each layer of the model.
- target_modules = ["q_proj"] - This is a list of layers in the model that will be updated during fine-tuning. In this case, we only update the q_proj layer.
- lora_dropout = 0.05 - Determines the proportion of neurons to be disabled during training to prevent overfitting.
- task_type = "CAUSAL_LM" - This is the type of task for which we are training the model. In this case, we are training the model for text generation based on sequential prediction of the next token (causal language modeling).

6 Epochs testing

Before training the model, the language of a generated text was boring, primarily consisting of a list of actions performed by the main character. As the model was trained, the generated text became more engaging and nuanced, demonstrating the effectiveness of the chosen method.

After the first epoch of training, we obtained a model capable of generating more lively text based on the provided "plot" from a single sentence. The model was able to add new ideas to the provided plot and to provide the text with some special vibes that resembled a typical fairy tale or fantasy story:

Plot: "Harry Potter meets a magical spider. "

Epoch 1: "Harry Potter was walking through the forest when he stumbled upon a small, furry creature with eight legs. It was a spider, and it was very, very magical. <...> "Hello, Harry," said the spider. "I am a magical spider, and I can see into the future. <...>"

After the third epoch generated texts revealed to be written in a language rich in metaphors and complex vocabulary:

Epoch 3: "Harry Potter was walking through the forest when he suddenly felt a tickling sensation on his neck. He looked up and saw a small spider crawling up his hair. <...> It was a magical spider, and it had the power to grant wishes. <...> "

Epoch 4: "Harry Potter was walking through the forest one day when he came across a spider. It was a big, fat spider, with a glossy black body and a large, golden head. Harry was fascinated by the spider's appearance, and he was about to move on when he noticed that the spider was wearing a large, golden ring on its head. Harry was very curious about the ring, and he decided to investigate. <...> "

As the model continued to train, it became increasingly difficult for it to generate text based on a single, small sentence. In earlier epochs, the model excelled at "expanding the idea," while in later epochs, it focused on generating text strictly within the given "plot" scope, resulting in a more detailed storyline. However, when increasing the maximum number of tokens, later epochs began to duplicate the same idea presented in the plot throughout the text.

Text 4: "<...>Harry had never heard a spider speak before. He was afraid to say anything. "I am not a

spider," he said. "I am a boy." "I know that," the spider said. "But I am a spider. I want to be your friend." Harry was still very surprised. He did not know what to say. "I am not a boy," he said again. "I am a man." "I know that," the spider said. "But I am a spider. I want to be your friend."

7 Evaluation

7.1 Automatic evaluation

The final text generated by the fine-tuned Mistral-7B-Instruct-v0.2-GPTQ was evaluated with GAPELMAPER Metric (Mikhaylovskiy, 2023; Mikhaylovskiy and Churilov, 2023). The evaluation results of the generated text are presented in Table 1, alongside the results calculated for well-known books' ("Don Quixote" and "The Adventures of Tom Sawyer") and one generated by S4 text's scores taken from Mikhaylovskiy and Churilov (2023). Mikhaylovskiy (2023) hypothesize that "GAPELMAPER less than 1 means that the autocorrelations decay according to a power law and the text is structured in a way. GAPELMAPER more than 1 means that the autocorrelations decay according to an exponential law and the text is unstructured". According to this statement the resulting text of our fine-tuned model exhibits a structured composition. Furthermore, the metrics obtained are similar to those achieved by "Don Quixote".

	Power law MAPE	**Exp law MAPE**	**GAPEL-MAPER**
Don Quixote	0.20	0.44	0.45
The Adventures of Tom Sawyer	0.21	0.55	0.38
S4 generated text	0.21	0.5	0.38
Mistral-7B-Instruct-v0.2-GPTQ Fine-tuned	0.17	0.402	0.44

Table 1: Automatic Evaluation Results

7.2 Human evaluation

The resulting fanfic was also subjected to human evaluation, as part of a blind assessment conducted by linguistics students. The participants were presented with a selection of anonymous fanfics, some of which were genuine and others generated by neural networks, without knowledge of their origin. After reading the texts, they completed a questionnaire, the results of which are presented in Table 2. Based on these scores, it can be inferred that the generated text demonstrates a satisfactory level of coherence, although articulating its core idea proves somewhat challenging.

From a language perspective, the text had some repetition, which suggested a limited vocabulary. However, the text's sentence structure was more complex, with features like parenthetical phrases, subordinate clauses, and participial phrases. The text also used common metaphors, comparisons, and oxymorons, but didn't go beyond these familiar expressions. The chapters showed a high degree of narrative repetition, leading one evaluator to suggest that the text was not written by a human. However, if this repetition is set aside, the evaluator believed that the writing style was typical of a young adult author who is a fan of the Harry Potter books.

Metric	**Score**
Correlation between the fanfic title and its content	3.25
Compatibility of chapter and sub-chapter titles with the overall style of the text	3.2
The strength of the stylistic connection between all the elements of the text	2.6
The pace of the plot	1.8
Word repetitions	2.6
Text composition	2.8
General idea of the text	3.2

Table 2: Artistic Quality Assessment Results

8. Results

As a result of our extensive research and experimentation, we were able to successfully fine-tune the model, testing it at various training epochs and gathering valuable insights into the aspects that require attention when fine-tuning it. One of the key findings from our investigation was that our fine-tuned model currently lacks the ability to make smooth transitions from one piece of text to

another, specifically from one chapter to another. This limitation resulted in the appearance of sudden and unexpected plot twists, as well as the repetition of similar scenarios in adjacent chapters, which was readily apparent to our informants.

Despite this limitation, we were also able to generate a long and cohesive artistic text, which exhibited a certain level of structural quality, as measured by the results of our metrics (score of 0.44). Notably, this text featured a logical beginning and end, demonstrating a clear narrative arc. This achievement is significant, as it suggests that our fine-tuned model is capable of producing texts that are not only coherent but also engaging and well-structured.

Our research highlights the importance of addressing the issue of smooth transitions between chapters, as this is a critical aspect of creating a compelling and immersive narrative. By refining our model to better handle these transitions, we can potentially improve the overall quality and coherence of the generated texts.

Acknowledgements

The authors are grateful to their colleagues at NTR Labs ML division and classmates at HSE for the discussions and support. Our earlier discussions with Nikolay Mikhaylovskiy enabled us to successfully implement our idea and achieve our current result. We are also extremely grateful to Anastasia Kolmogorova for insightful discussions and valuable guidance throughout this process.

References

Jian Guan, Zhuoer Feng, Yamei Chen, Ruilin He, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2022. *LOT: A Story-Centric Benchmark for Evaluating Chinese Long Text Understanding and Generation.* Transactions of the Association for Computational Linguistics, 10:434–451.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. *Mixtral of Experts*. arXiv: 2401.04088

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, William El Sayed. 2023. *"Mistral 7b",* arXiv: 2310.06825

Yukyung Lee, Soonwon Ka, Bokyung Son, Pilsung Kang, Jaewook Kang. 2024. *Navigating the Path of Writing: Outline-guided Text Generation with Large Language Models,* Korea University, Naver AX Center, arXiv:2404.13919v1

Aleksandr Migal, Daria Seredina, Lyudmila Telnina, Nikita Nazarov, Anastasia Kolmogorova, Nikolay Mikhaylovskiy. *Overview of Long Story Generation Challenge (LSGC) at INLG 2024*. In Proceedings of the 17th International Natural Language Generation Conference, Tokyo, Japan

Nikolay Mikhaylovskiy. 2023. *Long Story Generation Challenge*. In Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges, pages 10–16, Prague, Czechia. Association for Computational Linguistics.

Nikolay Mikhaylovskiy and Ilya Churilov. 2023. *Autocorrelations Decay in Texts and Applicability Limits of Language Models*. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2023"

Project Gutenberg. (n.d.). Retrieved February 21, 2024, from www.gutenberg.org.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, et al. 2021. *Multitask Prompted Training Enables Zero-Shot Task Generalization.* ICLR 2022.

Xiaofei Sun, Zijun Sun, Yuxian Meng, Jiwei Li, Chun Fan. 2022. *Summarize, Outline, and Elaborate: Long-Text Generation via Hierarchical Supervision from Extractive Summaries.*, arXiv:2010.07074v2