

Fine-Tuning a LLaMA Model for Spoiler-Free Book Summarization

Lathan W. Gregg
University of Virginia
uua9gw@virginia.edu

Kurt L. C. Samuels
University of Virginia
nps3cs@virginia.edu

Zach E. Stautzenbach
University of Virginia
ehe5bn@virginia.edu

Abstract

In this paper, we attempted to fine-tune a large language model to answer detailed, book-specific questions based on a reader’s current location within the text, without revealing future events and spoiling key plot points. We decided to use the first five books in the Harry Potter series and LLMs generated prompt-response pairs as our data for fine-tuning. A 7 billion-parameter LLaMA model from Hugging Face was first fine-tuned on the full Harry Potter text, and then on a large number of spoiler-free question-answer pairs containing information from various points within each chapter of the series generated using OpenAI’s GPT-4o model. 50 additional prompts were then created to evaluate the model’s performance. After fine-tuning, the generated text was evaluated with cosine similarity and human feedback by comparing the model output to the optimal response of our test prompts. Ultimately, model performance increased after each fine-tuning phase and generalized well against unseen questions, but the spoiler-free nature was occasionally broken.

1 Introduction

Consider the following scenario: you just decided to resume reading a book that you haven’t touched in weeks. You don’t want to start over but at the same time, you want a refresher on what you have already read without information about future events that may spoil the book.

A great solution would be to have a large language model (LLM) that can respond to questions you have about your book based on where you stopped reading so that you can resume where you left off without forgetting any key parts. This is the motivation behind our LLM project, to fine-tune a LLaMA model to be able to carry out spoiler-free book summarization for readers to be able to get a recap on what they have read or what they might have missed. In our case, we experimented with the Harry Potter book series.

2 Related Work

2.1 Using LLMs for Summarization

Using LLMs to summarize certain key pieces of information of large documents is very beneficial. This process is not the same as generating summaries of the whole document. This key idea is the difference between aspect based summarization, which is targeting key ideas or topics within a larger document, and generic summarization which is highlighted in the paper, “Leveraging the Power of LLMs: A Fine-Tuning Approach for High-Quality Aspect-Based Summarization” [1].

A key question this paper looked to answer is: does fine-tuning LLMs provide a significant benefit for aspect-based summarization tasks? This relates to our work, which aims to summarize relevant aspects of a book series based on user questions. The data experimented on in the paper was a domain and aspect set, where domain refers to varying topics such as education, music,

*Course Project for DS-6051 – Decoding Large Language Models.

career, and more. The authors wanted to be able to prompt a fine-tuned LLM to answer questions relating only to a certain domain or aspect set. Similar to our approach, a LLaMA model was fine-tuned for their task. Quantized Low-Rank Adaptation (QLoRA) and Parameter-Efficient Fine-Tuning (PEFT) was used to amplify the fine-tuning process. The authors evaluated their model’s output using metrics including Rouge 1 to evaluate the quality of generated summaries and by using GPT-4o as a critic. Both of these metrics were used to check how well the model’s responses were compared to the gold standard. More specifically, their model’s responses were judged based on relevance, coverage, impurity, rating, and goodness which allowed for them to measure the quality of the summary and the similarity. This paper concluded with the fine-tuned LLaMA-13b-FT model being optimal for this type of task. It was able to successfully identify and generate specific summaries of what a user specified rather than a generic summary of the entire document.

2.2 Utilization of Parameter-Efficient Fine-Tuning

Fine-tuning LLMs can be very computationally expensive. We researched multiple strategies to speed up this process while retaining model quality. In Weyssow et al, the authors suggest PEFT to do this [2]. This paper demonstrates its ability to retain model performance by training only a subset of the model parameters. Low-rank adaptation (LoRA) is a common type of PEFT which introduces low-rank adapters to the attention projections leading to drastically reduced computation time. For our project, we used LoRA to train adapters that were a fraction of the size of our full LLaMA model. The authors benchmarked various pre-trained coding models with different fine-tuning techniques such as full fine-tuning, LoRA, prompt tuning, IA3, and prefix tuning. The findings of their experiment concluded LoRA resulted in the best performing LLMs. These findings led us to apply this method on the LLaMA-2-7B-32K-Instruct pre-trained model.

3 Methodology

3.1 Data

The data used for fine-tuning our LLM consisted of the first five Harry Potter books:

1. Harry Potter and the Sorcerer’s Stone
2. Harry Potter and the Chamber of Secrets
3. Harry Potter and the Prisoner of Azkaban
4. Harry Potter and the Goblet of Fire
5. Harry Potter and the Order of the Phoenix

Preparing the data to be used in the fine-tuning process required a standardized format. First, all the PDF files of the books were converted to text files. Next, these text files were cleaned of unrelated text and Unicode characters. Once all text files were compiled, we transformed the data into a JSON file with book location metadata including the title of the book, the chapter of the book, and the associated text. The format of this file is shown in Figure 1 below.

```
{
  "index": 0,
  "book": "Harry Potter Book Title",
  "chapter": "Chapter Title",
  "page": #,
  "text": "Associated Text,"
}
```

Figure 1: Book Meta-data File Format

In addition to the structured book text, we also used GPT-4o to generate around 900 question and answer (QA) pairs to fine-tune the model on certain types of prompts. These consisted of a wide variety of topics such as characters, settings, themes, and symbolism. These also were stored within a JSON file containing an instruction referencing the book and chapter title and the GPT-4o generated answer. Figure 2 below shows the format of these QA pairs.

The final piece of data we had was an additional 50 QA pairs (10 per book) to test and evaluate how our fine-tuned model performs on unseen questions.

```
{
  "Instruction": "You are reading  
Harry Potter X, Chapter Y,  
Question: what happens?",
  "Response": "This happens in Chapter  
Y."
}
```

Figure 2: Question and Answer Data Format

3.2 Model Architecture

We selected the LLaMA-2-7B-32K-Instruct model as the foundation for spoiler-free book summarization. This model is a fine-tuned version of Meta’s LLaMA 2 architecture, trained for human instruction and long-context summarization.

We employed PEFT using LoRA to retain the knowledge of the original model and to adhere to the computational limitations of our training environment. Rather than training all 7 billion parameters, LoRA introduces low-rank matrices into attention projections, significantly reducing the memory and computational requirements while maintaining the quality of the model.

Our configuration used a LoRA rank of 8, a scaling factor of 16, and a dropout rate of 0.05, resulting in approximately 4.2 million additional trainable parameters, which is less than 0.1% of the model’s total size. Mixed-precision (16-bit) training and gradient checkpointing were also utilized to minimize memory usage, allowing fine-tuning on a single NVIDIA A100 GPU with 40 GB of memory.

3.3 Fine-Tuning Procedure

We fine-tuned the model in two sequential stages to incrementally build spoiler-free book summarization capability. In the first stage, the model was trained on the structured book text. In the second stage, we used the QA pairs to fine-tune the model to encourage spoiler-free reasoning and demonstrate the desired response behavior. This stage aimed to reinforce the model’s ability to provide concise, relevant responses without revealing future developments.

Training was conducted with a batch size of 1, gradient accumulation over 16 steps, and a learn-

ing rate of 0.0002. Each stage was trained for 6 epochs. This multi-stage fine-tuning was designed to allow the model to first acquire a deep understanding of the book context and then refine its ability to respond to prompts accurately while avoiding spoilers.

3.4 Evaluation Metrics

We evaluated the model’s performance on 50 QA prompts withheld from the training process. Evaluation was done at multiple stages using qualitative and quantitative methods to assess the model’s ability to generate relevant, spoiler-free responses.

3.4.1 Cosine Similarity

To measure the semantic similarity between generated responses and the gold standard responses, we computed cosine similarity scores. Embeddings for both were obtained using the all-MiniLM-L6-v2 Sentence-BERT model. Cosine similarity scores were computed by measuring the cosine of the angles between the vectors in the embedding space. Values range from -1 to 1 , with higher values indicating high semantic similarity between the model’s response and the gold standard. Importantly, cosine similarity is able to evaluate the relevance of the model’s response, even when the responses are phrased differently than the expected answers.

3.4.2 Spoiler-Free Accuracy

In addition to the cosine similarity, we also manually reviewed responses from the test questions generated by the model trained with only the text and the model fine-tuned using the two stage approach. For each response, we first evaluated if the model provided a coherent response aligned with the desired summarization and response behavior. For coherent responses, we identified if the responses were accurate to the events of the book. Finally, for accurate and coherent responses, we labeled the responses based on their accuracy in avoiding spoilers.

3.4.3 Qualitative Accuracy

We also conducted qualitative analysis by comparing the model generated responses to the gold standard and our own knowledge of the books. Particular attention was paid to spoiler leaks, hallucinated details, and speculative answers.

Together, these evaluation metrics provided a comprehensive understanding of the model’s capacity to provide highly relevant and spoiler-free responses, and how capabilities evolved at each stage of the training process.

4 Experiments and Results

4.1 Training Performance

Using a single NVIDIA A100 GPU, the model achieved a training time of 17 minutes and 50 seconds for text-only fine-tuning and 59 minutes and 29 seconds when fine-tuned on question-answer pairs. The final cross-entropy loss was 2.014 for the text-only model and 0.662 for the fully fine-tuned model.

4.2 Quantitative Evaluation

Table 1 summarizes the average cosine similarity and standard deviation across three stages of training: the base model without fine-tuning, the model fine-tuned only on structured book text, and the model fine-tuned sequentially on both book text and synthetic question-answer pairs.

The model fine-tuned on both book text and question-answer prompts achieved the highest average cosine similarity score of 0.57, outperforming both the base model (0.52) and the model fine-tuned solely on text (0.53). This indicates that the two-stage training process improved the ability of the model to generate responses that align with the desired responses.

The standard deviation of the final model was slightly lower than the base model and the model trained on only the book text, indicating that improvements were due to a general increase in quality, rather than a few isolated examples.

Table 2 contains the labeled responses for the model fine-tuned on only the text and the model fine-tuned on the text and QA pairs.

The model fine-tuned solely on structured text demonstrated very limited coherence and factual accuracy, with only 3 out of 50 responses judged to be coherent and none judged to be accurate. In contrast, the model fine-tuned on both text and QA pairs produced coherent outputs in 48 out of 50 cases, accurate responses in 24 cases, and spoiler-free answers in 22 cases.

Train Stage	Avg. Cos. Sim.	Std. Dev.
Base Model	0.52	0.14
Text-only	0.53	0.14
Text + QA	0.57	0.12

Table 1: Cosine similarity scores comparing generated responses to gold answers across different fine-tuning stages.

Train Stage	Coherent	Acc.	Spoiler-Free
Text-only	3	0	0
Text + QA	48	24	22

Table 2: Manual evaluation of model outputs for coherence, accuracy, and spoiler-freeness across two model configurations.

4.3 Qualitative Analysis

In addition to quantitative analysis, we also compared the responses generated by the model at each stage to understand their strengths and limitations. Multiple examples of model outputs to test prompts are provided in the appendix Table 3.

Overall, the base model tended to provide responses that were irrelevant and did not align with the desired response formatting. For many prompts, the model proposed new questions rather than answering the provided question or otherwise provided incorrect or hallucinated responses.

While, the model fine-tuned on just the structured text tended to provide responses featuring correct character names and book locations, it struggled to provide coherent and accurate responses. The model frequently generated hallucinated excerpts and similar to the base model,

often generated new questions rather than an answer to the prompt.

The model trained on both the text and QA pairs demonstrated a substantially improved ability to provide answers that modeled the desired behavior. The responses were typically concise and relevant to the question prompt. The responses consistently referenced specific characters and details from the book but often confused nuanced details.

These examples highlight key qualitative improvements:

- **Spoiler Avoidance:** Fine-tuned models better constrained their answers to information revealed up to the current chapter.
- **Contextual Precision:** Fine-tuned models framed details appropriately without unnecessary speculation or foreshadowing.
- **Factual Accuracy:** Fine-tuned models were less likely to hallucinate facts not present in the current context.

5 Discussion

5.1 Successes and Strengths

Throughout this process, we successfully fine-tuned the model to get closer to reaching our goal of spoiler-free book summarization at each iteration. The model saw improved performance as more context was given.

First, when we fine-tuned on text only, the model was able to output somewhat relevant information that related to Harry Potter. It had a lot more knowledge on the first five books and could answer more about Harry Potter than the base model. However, it was not very coherent and often outputted hallucinated text.

Next, when we used our generated prompt and response pairs to instruct the model how to respond to certain questions, we saw a significant improvement. By prompting the model what book and chapter the reader is on, the model could understand a little bit more the general location of where to get information. This led to nearly all of the test prompts, 48 out of 50,

being coherent, with around half answering the questions correctly. Of the questions that were correctly answered, all but two were spoiler-free. This stage of fine-tuning was a big step forward in that our model became more concise and would respond to prompts with more relevant information. It also transitioned from repeating the text to actively generating more thoughtful answers. It demonstrated the ability to generalize effectively to unseen questions.

5.2 Challenges and Limitations

While the two-stage fine-tuning approach significantly improved the model’s ability to generate coherent and spoiler-free responses, several limitations were observed throughout the project.

First, despite improvements, the fine-tuned models consistently struggled with nuanced or subtle details from the books, generating accurate responses for less than half of the test prompts. This was especially pronounced when the correct answer required inference rather than direct retrieval. Modifications to the Stage 1 training setup, such as incorporating additional annotations, highlighting key events, or supplementing the book text with high quality summaries, could enhance the model’s ability to grasp more detailed narrative elements.

Second, the performance of the base model was significantly worse than expected for a fine-tuned version of a successfully and widely used LLM. We believe additional hyperparameter tuning for the base model, such as adjustments to the temperature and maximum number of generated tokens, could have improved the baseline model’s performance, particularly in producing more coherent outputs. A more systematic search for optimal configurations may have led to stronger baseline comparisons and final model performance.

Finally, this project did not incorporate prompt engineering, few-shot learning, or reinforcement learning with human feedback (RLHF) to enhance the model’s spoiler-free generation behavior. These techniques could potentially guide the model to a more nuanced understanding of the task and further improve the

output formatting and spoiler-avoidance without relying solely on supervised fine-tuning.

6 Conclusion and Future Work

6.1 Conclusion

In conclusion, the fully fine-tuned model greatly outperformed the base and text-only models. The base model frequently provided unrelated information and often formatted responses incorrectly. The text-only model demonstrated an improved ability to reference specific details from the book, but struggled to provide coherent and accurate responses. It frequently quoted the text, and sometimes produced completely unreadable hallucinations. The final model, fine-tuned on the structured text and QA pairs resulted in coherent responses that were entirely spoiler-free.

6.2 Future Work

An important next step in our future work will be training a reward model for spoiler control by incorporating RLHF. By training a reward model on spoiler-free human-preferred responses and assigning a higher score to spoiler-free outputs, our fine-tuned model will produce a robust and detailed response within a spoiler-free context.

Prompt engineering and in-context learning are techniques that would further enhance the behavior of the model. Carefully crafted prompts or few-shot prompting could allow the model to become more context-aware without needing to repeat the full fine-tuning process. Combining a reward-based model with the revised prompt-based methods could yield improvement to the model’s accuracy, relevance and spoiler avoidance.

An additional extension of our research would be scaling our dataset to a wider array of novels. This would allow our fine-tuned model to generalize its spoiler awareness over various genres, literary styles, and plot structures. However, this comes with the challenge of creating significantly more labeled prompts for each subsequent

book.

References

- [1] A. Mullick, S. Bose, R. Saha, A. K. Bhowmick, A. Vempaty, P. Goyal, N. Ganguly, P. Dey, and R. Kokku. Leveraging the power of llms: A fine-tuning approach for high-quality aspect-based summarization. <https://arxiv.org/html/2408.02584v1#S2>, 2024. arXiv preprint.
- [2] M. Weyssow, X. Zhou, K. Kim, D. Lo, and H. Sahraoui. Exploring parameter-efficient fine-tuning techniques for code generation with large language models. <https://arxiv.org/pdf/2308.10462>, 2024. arXiv preprint.

A Appendix

Prompt	Golden Response	Base Model	Text-only	Text+QA
<i>You are reading *Harry Potter and the Sorcerer's Stone*, Chapter 7: "The Sorting Hat". Question: Why does the Sorting Hat strongly consider Slytherin for Harry?</i>	"It sees ambition, cleverness, and a thirst to prove himself — traits of Slytherin. But Harry's refusal to join Slytherin influences the final decision, emphasizing the power of personal choice."	"Because he has a lot of Slytherin traits. (book, chapter 7, page 124) But Harry has never wanted to be a member of the Slytherin Quidditch team. (book, chapter 7, page 124)"The Sorting Hat knows," said Dumbledore. "The Sorting Hat sees more than we do." "But what if the Sorting Hat put Harry in Slytherin?" asked Hermione. "He's got a lot of Slytherin traits."	"(Slytherin was the worst house, the Sorting Hat explained.) "You've got the worst possible birthday," said the Sorting Hat, "you're a Slytherin and a Potter, you're the most famous wizard family there is, so they'll all be watching you. If you're not careful, they'll be all over you like a rash. But there's another reason they'll be watching you, because you're the only wizard in the world who's got a chance of beating the Sorcerer's Stone."	"Answer: Harry shows strong ambition and inner strength, which the hat sees as Slytherin traits. Ultimately, he's placed in Gryffindor. This happens in Chapter 7. This moment is crucial to Harry's character and sets up his values. It's described in Chapter 7."
<i>You are reading *Harry Potter and the Chamber of Secrets*, Chapter 2: "Dobby's Warning". Question: What is the effect of Dobby using magic in the Dursleys' house?</i>	"Harry receives a warning from the Ministry of Magic, which doesn't distinguish who cast the magic, putting him in further trouble. This occurs in Chapter 2"	"Dobby is expelled from Hogwarts (Document[170]):"Dobby makes his way to Privet Drive and performs his duty, despite the fact that he is expelled from Hogwarts for doing so . Dobby appears in the Dursleys' house, and warns Harry Potter that the Dursleys are in grave danger . Dobby uses magic to alert Harry to the Dursleys' presence, and Harry wakes up to find Dobby standing in front of him."nDobby uses magic in the Dursleys' house"	"(G) Dobby had been right, Harry's aunt was in a terrible temper that morning. She shouted at Uncle Vernon for making her late for work and she was still cross when Uncle Vernon sent Harry out to get the mail, because it was a rainy morning. Uncle Vernon had to go into the post office to collect the mail for the whole street and Harry had to wait under the front door in the pouring rain. He was soaked through in no time. When Uncle Vernon returned, he was still angry. "It's all this bloody muggle's fault!"	"Answer: Dobby's use of magic causes the Dursleys' house to shake, and the Dursleys become aware of his presence. This incident occurs in Chapter 2. It shows Dobby's desperation to help Harry."

Table 3: Example responses comparing the base model and fine-tuned model outputs (continued in 4

Prompt	Golden Response	Base Model	Text-only	Text+QA
<i>You are reading *Harry Potter and the Sorcerer's Stone*, Chapter 17: "The Man with Two Faces".: What causes Quirrell's body to fail when he tries to take the Stone from Harry?</i>	Harry's mother's sacrificial protection makes it physically painful for Quirrell to touch him. This enchantment is what ultimately stops Quirrell and protects the Stone.	A) Harry uses magic to defend himself) Quirrell's body has been weakened by the Stone) Voldemort takes control of Quirrell's body) Harry Potter's magic is stronger than Quirrell's The correct answer is B) Quirrell's body has been weakened by the Stone (Document[17]):"Harry Potter and Dumbledore return to Hogwarts, where Dumbledore explains that the Stone is powerless without its container . Voldemort, who has taken control of Quirrell,	(Dumbledore) Harry: "It's all right, it's all right, it's all right. I've got it. I've got it. It's mine. I've got it. It's mine." The other two men were staring at Harry in amazement. "It's all right, it's all right, it's all right. I've got it. I've got it. It's mine. I've got it." "What's wrong with you?" "It's all right, it's all	Answer: Harry's emotional connection to the Stone is stronger than Quirrell's intellectual knowledge, creating a magical resistance. This emotional barrier is explained in Chapter 17.nSummary: Chapter 17.

Table 4: (continued) Example responses comparing the base model and fine-tuned model outputs continued.