Novelty: Optimizing StreamingLLM for Novel Plot Generation

Stanford CS224N Custom Project

Joyce Chen
Department of Data Science
Stanford University
joycech@stanford.edu

Megan Mou
Department of Computer Science
Stanford University
meganmou@stanford.edu

Abstract

Large Language Models (LLMs) struggle with the task of long-form creative writing both technically, by exhibiting undesired behavior when the context length exceeds its training sequence length, and creatively, by producing bland plots that do not develop a coherent or engaging story. While very large models with long context lengths such as GPT-4 have made strides in this area, we aim to optimize for creativity and diversity of natural language generation with computational constraints and a small underlying context window of 4k tokens.

We introduce the novel approach of treating novel-writing as a streaming "question-answer" chapter-generating conversation between user and assistant. We produce perceptible improvements in model creativity by building a rich novel plot generation model through fine-tuning Vicuna-7b with StreamingLLM on chapter-by-chapter detailed plot summaries. We also demonstrate that different training data formats and prompt engineering at inference time can produce greater diversity of output and help prevent repetitive rabbit-holing.

1 Key Information

- Mentor: Tianyi Zhang
- Contributions: Megan scraped and processed all of the datasets for training and testing. Joyce fine-tuned the models and ran experiments. Megan and Joyce worked equally on evaluation code and writing.

2 Introduction

Novel plot generation presents a challenge both for long sequence coherence and model creativity. A good novel-generating model must keep track of characters and their relationships, while also prudently judging when to introduce new plot points, plot twist, and characters. Outside of the novel-generating task, simply *extending* the meaningful output of a model beyond its pretraining sequence length is difficult – we are adding the additional challenge of *developing* upon chapters that are no longer in context.

It is clear from related work that a larger context window will result in significant performance boosts for the task of coherent long-text generation (Peng et al., 2023). However, longer context models require more memory and processing power, which we and many others do not have access to. Furthermore, since reducing hallucination is a priority for researchers building foundation models, we find that the creative capabilities of smaller open-source models are often lacking beyond generating a limerick or two.

In order to address the issue of extending coherence beyond the pretraining sequence length, we utilize StreamingLLM, a very lightweight addition that is compatible with many open-source LLMs (Xiao et al., 2023). To improve model creativity for the task of writing a novel, we fine-tune Vicuna-7b-v1.5 on hundreds of high-quality novel chapter summaries. Finally, we overlay StreamingLLM's attention sink and rolling key-value cache mechanism on top of our fine-tuned Vicuna models at inference time.

Combining these approaches, we build a model that achieves better performance on the length, diversity, and coherence of generated novel plots with limited computational resources – and produce some pretty interesting stories along the way.

3 Related Work

StreamingLLM exists in the broader problem space of improving long human/machine assistant language interactions. Research in this domain falls primarily under three main categories:

1. **Length extrapolation**. This aims to improve LLM performance on longer texts at test time after only being trained on shorter sequences.
2. **Context window extension**. This aims to extend models' dense attention windows in order to capture longer surrounding context when processing tokens during forward pass.
3. **Improving LLMs' utilization of long text**. This aims to more effectively utilize content that is already present in LLM context windows; i.e. making more sense from the context than simply using it as input.

It is important to note that StreamingLLM only addresses the first of these areas — the streaming mechanism helps model performance generalize from finite length attention windows during pre-training to infinite sequence lengths at inference time.

One research area within the category of context window extension is positional interpolation. Recent work includes YaRN (Peng et al., 2023)], which extends the context windows of LLMs pretrained with RoPE (Rotary Positional Embeddings) by scaling positional embeddings based on the context length. Another example of work in this domain is fine-tuning using LongLoRA, which effectively extends the context sizes of pre-trained LLMs through shifted sparse attention at a lower computational cost (Chen et al., 2023).

We were also inspired by other work in the realm of novel-generation. These primarily employ inference-time methods to increase coherence; for example, generating longer stories with recursive re-prompting and revision during inference (Yang et al., 2022b) or outline-controller models that prompt the model to generate an initial plot outline for stories and enforces subsequent plot generation to adhere to that outline via a unique discriminator (Yang et al., 2022a).

4 Approach

We used Vicuna-7b-v1.5 as our base model, which is LLaMA-2 (a transformer-based open-source LLM) fine-tuned on 125k instruction-following conversations from ShareGPT (Zheng et al., 2023a). Vicuna is a very popular open-source model that is commonly used as the base model for open-source research; we combine it with StreamingLLM for our streaming baseline. We LoRA fine-tuned this base model with 4-bit quantization on our two instruction-following datasets; the smaller dataset consisting of 22 novels, and the larger dataset consisting of 45 novels (988 novel chapters). LoRA stands for low-rank adaptation; the method freezes the pre-trained model weights, instead training rank decomposition matrices (Dettmers et al., 2023). LoRA vastly reduces the number of trainable parameters with comparable results as full fine-tuning, and is therefore used very widely to reduce the computational requirements for fine-tuning.

Next, we applied StreamingLLM's attention sink and rolling key-value (KV) cache mechanism on top of the fine-tuned model; this works at inference time. The attention sinks are four initial tokens strategically chosen to address a key feature of autoregressive language models. In such models, a significant portion of the attention distribution is allocated to the initial tokens of an input sequence, as evidenced by the e^{z_1} term, where $z_1 >> z_j$ in the denominator of the softmax equation (1) below. Adding these starting tokens' KV to the attention computation of a current sliding window keeps the attention distribution close to normal, stabilizing performance. This is because removing the e^{z_1} term

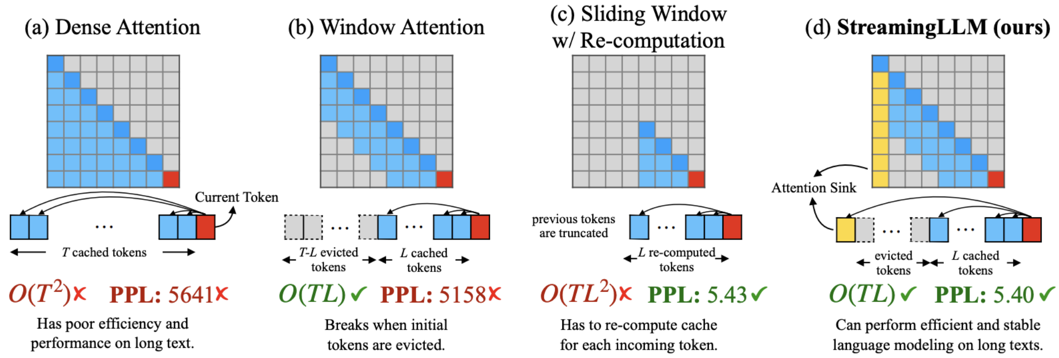
(a) Dense Attention
(b) Window Attention
(c) Sliding Window w/ Re-computation
(d) **StreamingLLM (ours)**
Current Token
T cached tokens
$O(T^2)$ ✗ PPL: 5641 ✗
Has poor efficiency and performance on long text.
T-L evicted tokens
L cached tokens
$O(TL)$ ✓ PPL: 5158 ✗
Breaks when initial tokens are evicted.
previous tokens are truncated
L re-computed tokens
$O(TL^2)$ ✗ PPL: 5.43 ✓
Has to re-compute cache for each incoming token.
Attention Sink
evicted tokens
L cached tokens
$O(TL)$ ✓ PPL: 5.40 ✓
Can perform efficient and stable language modeling on long texts.

Figure 1: Compared to other attention methods, StreamingLLM keeps the attention sink (several initial tokens), combined with recent tokens in a rolling KV cache (Xiao et al., 2023)

from the denominator would decrease the denominator substantially, so retaining these initial few tokens improves model performance on significantly longer sequences (Xiao et al., 2023).

$$\sigma(z_i) = \frac{e^{z_i}}{e^{z_1} + \sum_{j=1}^{K} e^{z_j}} \quad for\ z_1 >> z_j, j = 2, \ldots, K \tag{1}$$

The most recent tokens are stored in the rolling KV cache, where the positional information assigned to the recent tokens are relative to their positions within the cache, rather than in the original corpus itself. This rolling cache is similar to previous sliding window attention implementations.

StreamingLLM allows Vicuna to generalize beyond its pre-trained sequence length in order to continue generating coherent text for (theoretically, as we will discuss in our results) millions of tokens. We set the start size and recent size of StreamingLLM to 4 and 4000 respectively, with the effect of storing the 4 initial tokens' and 4000 most recent tokens' key-value pairs in the cache.

We wrote novel code to run StreamingLLM with our fine-tuned Vicuna models since StreamingLLM was not built to operate with fine-tuned models. Additionally, we wrote code to allow for dynamic inference, where the user can enter a prompt from the command-line and receive a response (as opposed to inference on pre-made test scripts). Inspired by other studies that also attempted to mitigate repetition in neural generation (Xu et al., 2022), we wrote code to perform n-gram blocking at inference time, as well as token-by-token cache eviction and input; however, these experiments resulted in extremely slow and poor performance.

Another key method involved extensive iteration on our instruction-following data format for both the training and test datasets. We designed several different versions of the prompts for our training data by evaluating the quality of response from ChatGPT for each potential prompt. We also tried a dozen different inference time prompts (see A.3).

We ended up fine-tuning on two types of training prompt format:

1. Provide the plot summary of the previous chapter as context in every prompt. Prompt the model to generate the next chapter based on the previous chapter. We will refer to this format as [previous chapter in prompt].

2. Provide the plot summary of the first chapter of the novel in the first user prompt, then prompting the model to generate the next chapter without the context of the previous chapter in subsequent turns. We will refer to this format as [contextless prompt].

5 Experiments

5.1 Data

For our initial high-quality train and test datasets, we used "BOOKSUM: A Collection of Datasets for Long-form Narrative Summarization"(Kryściński et al., 2022a). We built upon and debugged their scripts for scraping novel chapter summaries from a variety of websites, including CliffsNotes and SparkNotes, which were the most detailed sources of chapter-by-chapter novel summaries. We scraped 22 chapter-by-chapter novel summaries for our initial smaller fine-tuning dataset.

To compile the larger dataset of 45 novels, we loaded in a HuggingFace dataset created using the same BOOKSUM dataset from above (all high-quality chapter-by-chapter summaries from CliffsNotes and SparkNotes) (Kryściński et al., 2022b). These 45 texts were majority fiction novels, with a few plays and nonfiction biographies. We manually assessed detail and coherence of the chapter summaries for each novel we picked for training, prioritizing high quality over high quantity data(Zhou et al., 2023).

We wrote our own scripts to convert the train/test splits of the data respectively into 1) an instruction-following question-answer format used to LoRA fine-tune LLaMA-2 (FastChat format (Zheng et al., 2023b)) and 2) the format of test examples used by StreamingLLM (Xiao et al., 2023), tailored for our task of sequential chapter summary generation (see A.1 and A.2). Each new chapter prompt is represented as a sequential turn within a "conversation" structure during both training and testing. Initially, the model is instructed to assume the role of an acclaimed author, tasked with generating one new following chapter based on the first chapter summary of a novel. In each subsequent turn of the conversation, the model is directed to produce the next new chapter, simulating a dynamic back-and-forth exchange between user and assistant.

5.2 Evaluation method

To holistically assess and compare the quality of chapter plots generated by our base and various fine-tuned streaming models, we employed qualitative and quantitative human evaluation. We collected this data via a Google Form survey, where we provided the second chapter of "Babbitt" as generated by each of our six distinct model/prompt combinations. Participants were asked to rate the quality of each generated summary on a scale from 1 to 10, without prior knowledge of which model had generated which summary. Additionally, we asked participants to specify why they found their top- and bottom-ranked plots most and least interesting respectively, in order to understand which qualities of the generated text impacted participants' impression of model performance.

We also measured model performance quantitatively by calculating the following reference-free scores across our five test novels:

1. Self-BLEU (SBL) to measure intra-novel diversity, i.e. the level of repetition between the chapters of each generated novel. Self-BLEU is calculated by taking the average BLEU score of a reference chapter against other chapters in the same novel. We take the average of the self-BLEU scores across all five generated novels for each model (Zhu et al., 2018).
2. Lexical Repetition-4 (LR-4) to measure intra-chapter diversity, i.e. how much repetition there is between sentences within a single chapter summary. LR-4 is calculated by taking the ratio of the count of 4-grams that appear more than once over the total number of 4-grams within a generated chapter. We calculate the average LR-4 score over all chapters generated by a model.
3. Distinct-4 (D-4) to measure inter-novel and overall diversity, i.e. the level of repetition across all generated novel plots. We do this by counting the percentage of distinct 4-grams (sequences of four words) across all content generated by the model (Li et al., 2015).

5.3 Experimental details

Fine-tuning on 22 novel dataset [previous chapter in prompt]

We first fine-tuned Vicuna-7b-v1.5 on our smaller high-quality dataset of 22 novels for 10 epochs, with an initial learning rate of 2e-5, a warm-up ratio of 0.03, and a cosine learning rate scheduler to adjust learning rates throughout training. We did not utilize a weight decay or gradient accumulation. We used a batch size of 2. Training time was approximately 20 minutes.

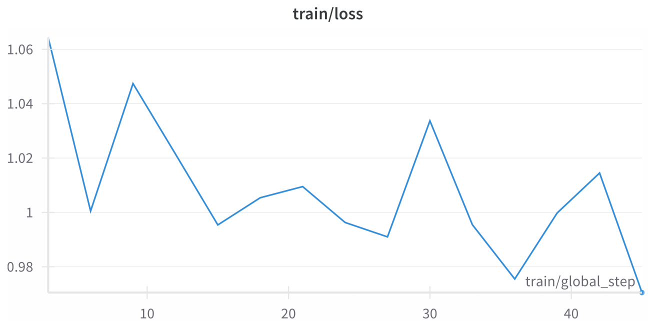

Figure 2: Training loss curve for smaller 22-novel dataset

Fine-tuning on 45 novel dataset [previous chapter in prompt]

We fine-tuned Vicuna-7b-v1.5 on our full dataset of 45 novels for 15 epochs. We kept most learning rate hyperparameters the same, but we included a weight decay of 0.001 to reduce the variance of the model and discourage over-fitting. After noticing that many training sequence lengths exceeded the model's default maximum sequence length of 2048, we increased the maximum sequence length to 2500 tokens. We used a batch size of 2 and set gradient accumulation steps to 4. This meant that we were accumulating gradients over 4 batches, each of size 2, before performing a single update to the model's parameters using these accumulated gradients. This allowed us to effectively use a larger batch size without requiring more memory. Training time was approximately 1 hour.

Fine-tuning on the 22 and 45 novel datasets [contextless prompt]

We altered both the 22 novel and 45 novel datasets so that the previous chapter was not provided in each human prompt in the novel "conversation"; we only provided the plot summary of the first chapter in each novel as the first turn of the conversation. We did this because we suspected that adding the previous chapter to the prompt could have been biasing the model towards repetition. We fine-tuned Vicuna on these new datasets using identical hyper-parameters as for the 22 novel dataset with previous chapter in prompt. Training on each dataset took approximately the same amount of time as training on the corresponding datasets with previous chapter prompts.

Fine-tuning on 998 individual chapter dataset [contextless prompt]

Finally, rather than treating each novel as a conversation consisting of (# of chapters * 2) question-answer (or prompt-chapter) pairs, we split up each novel so that every prompt-chapter pair would be treated as an individual conversation/training example. We fine-tuned Vicuna on this dataset for 10 epochs and otherwise kept the same hyperparameters as for the full-dataset fine-tuning. We used a batch size of 2 and set gradient accumulation steps to 4. Training time was around 6 hours.

We also tried to fine-tune the model with 8-bit quantization, but this failed with our computational resources. Therefore, our only option was to fine-tune with 4-bit quantization. The number of trainable parameters was 4,194,304, constituting 6.22% of all model parameters. We ran all training on 2 A4000 GPUs each with 16GB of memory.

5.4 Results

We ran each of our model/prompt combinations on our test dataset of 5 unique novels: Babbitt, Madame Bovary, The Red and the Black, Sense and Sensibility, and The Tempest. This means that there were five prompts with the context of the first chapter summary from that novel, and 45 prompts asking the model to generate the next chapter given its previous responses. This is a major challenge for all the models, as the context length of the ongoing "conversation" far exceeds the underlying Vicuna context length of only 4k tokens (Zheng et al., 2023a). This led many models to start repeating generated chapter plots, and all models apart from Vicuna fine-tuned on 22 novels with StreamingLLM ended up repeating sentences within a single chapter. We exclude the results from the model fine-tuned on 988 individual chapters as it failed from the first response inference, repeating the plot of the provided first chapter.

	Lexical Repetition (LR-4) ↓	Self-BLEU (SBL) ↓	Distinct-4 (D-4) ↑	Modal Plot Quality Score ↑
Base Vicuna-7b-v1.5 + StreamingLLM [B]	0.0496	0.6426	0.3435	6.5
Vicuna + StreamingLLM FT on 22 Novels w/ Prev Chapter [B]	0.2117	0.6133	0.2299	8
Vicuna + StreamingLLM FT on 22 Novels w/ Contextless Prompts [B]	0.1391	0.5945	0.3263	8
Vicuna + StreamingLLM FT on 45 Novels w/ Prev Chapter [B]	0.7008	0.8351	0.0137	9
Vicuna + StreamingLLM FT on 45 Novels w/ Contextless Prompts [B]	0.202	0.6026	0.1439	5.5
Vicuna + StreamingLLM FT on 45 Novels w/ Prev Chapter [D]	0.0523	0.0223	0.7528	9

Table 1: Comparison of un-fine-tuned Streaming Vicuna-7b vs. fine-tuned Streaming Vicuna-7b models. We tried two different test prompts: [B] Basic prompt, "Given your previous responses, which are plot summaries of the novel chapters you have written so far, write the detailed plot of the next single chapter of the novel." [D] Diverse prompt, 9 unique plot domain prompts (see A.2).

Model	**Excerpt from "Charles Bovary" 2nd Generated Chapter Summary**
Baseline Vicuna-7b-v1.5 + StreamingLLM [B]	The next chapter of the novel opens with Charles Bovary sitting in his office, staring at the ceiling. He is lost in thought, wondering if his life will ever be different. He has been a doctor for several years now, and while his practice has prospered, he is still unhappy. He feels trapped by his marriage to Heloise and the expectations that come with it.
Vicuna + StreamingLLM FT on 22 Novels w/ Prev Chapter [B]	Charles Bovary, now a successful doctor, is at home in his small village of Tostes. He is greeted by his wife, Heloise, who is eager to discuss the day's events. Charles is preoccupied with the recent death of a patient, a wealthy farmer named Homais. Homais was a close friend of Charles and had been instrumental in helping him establish his medical practice in Tostes.
Vicuna + StreamingLLM FT on 22 Novels w/ Contextless Prompts [B]	Charles Bovary, now a successful doctor, is at a crossroads in his life. He is torn between his love for his wife, Heloise, and his growing feelings for a new patient, Emma Bovary. As he spends more and more time with Emma, Charles becomes more and more consumed by his love for her. He begins to neglect his practice and his responsibilities, and his marriage to Heloise becomes more and more strained.
Vicuna + StreamingLLM FT on 45 Novels w/ Prev Chapter [B]	She was also clearly nervous, her hands shaking as she approached him. "Dr. Bovary, I'm in trouble," she said, her voice trembling. "I've been seeing a man, and I'm pregnant. I don't know what to do." Charles sighed, knowing all too well the struggles of young love. He had been there himself once, and he knew how difficult it could be.
Vicuna + StreamingLLM FT on 45 Novels w/ Contextless Prompts [B]	Charles Bovary, a shy and clumsy country lad, strikes his schoolmates as a diligent and industrious student. He is quiet and mixes well with the other boys. His father is a former army surgeon who had been forced to leave the service as a result of some scandal. He is a handsome and unscrupulous man who had married Charles' mother in order to get his hands on her large dowry. After the marriage, he wastes most of the money in foolish speculations, drinking, and amorous affairs.
Vicuna + StreamingLLM FT on 45 Novels w/ Prev Chapter [D]	Charles is torn between his own desires and the needs of those he cares about, and he struggles to find the right path. Meanwhile, his wife Heloise becomes increasingly jealous and possessive, and her behavior begins to strain their marriage. As Charles grapples with his decision, he is also forced to confront the reality of his own failures and shortcomings.

Table 2: Comparison of plain Vicuna-7b vs. Vicuna-7b with StreamingLLM enabled on 2 different versions of input prompts.

Our quantitative results are presented in Table 1, with the best LR-4, SBL, D-4, and plot quality scores across all models highlighted in green for each column. Overall, we observe that fine-tuning with contextless prompts in both the smaller and larger datasets had a slight positive impact on model performance compared to fine-tuning on datasets where the previous chapter was provided in the prompt. This is shown by the reduction of repetition and increase in diversity, as indicated by lower self-BLEU and lexical repetition scores, and a higher Distinct-4 score.

Since the human response data from our Google Form was very divided, with the 1-10 ratings for each model's performance ranging by about 8 points, we decided that the mode was the best average metric for the human-evaluated plot quality score. The variance in human-rated scores emphasizes the subjective nature of assessing quality fiction writing, and the difficulty inherent in performing and evaluating this task for both humans and language models alike. Nonetheless, results from the last column of Table 1 reveal that the highest plot quality scores were consistently obtained from models fine-tuned on 45 novels, confirming that our fine-tuning enhances the literary quality of model-generated content.

We were surprised to see that fine-tuning the model on all 45 novels resulted in more repetition at inference time for both training data types; we discuss this behavior more in Analysis. Overall, the model fine-tuned on 45 novels with *diverse* prompts at inference time consistently outperformed all other models across all metrics except for LR-4 where it closely trailed the baseline. This outcome was anticipated, since varying the prompts during test-time encourages the generation of more diverse content, which we will discuss further in our analysis on repetition.

6 Analysis

Repetition

From our qualitative human evaluation, it is clear that fine-tuning improves the quality, length and complexity of the generated chapter plots. However, after trying five different fine-tuned models (including the contextless prompts in training data), we found that fine-tuning cannot resolve the issue of model repetition across chapters. However, it can help vary sentence structure within a single generated chapter compared to the baseline un-fine-tuned Vicuna model with StreamingLLM.

The key characteristic of all of our models was the tendency to repeat previously generated content. In all cases, the model would repeat entire chapters in later turns within a novel; in some cases, the model would repeat sentences within a chapter to infinity and fail to terminate generation.

We found that all streaming models begin to repeat between chapters when the context window reaches just under 5k tokens, which matches the context length (4.7k tokens) where non-streaming Vicuna-7b-v1.5 begins to output gibberish in our initial experiments (Zheng et al., 2023a). We note that this characteristic of repetition is not exclusive to our task of longer-form novel chapter summary generation, but also occurs when running un-fine-tuned Vicuna + StreamingLLM on StreamingLLM's example prompts.

We hypothesize that the performance degradation of the streaming model occurs when the context length surpasses the pre-training context window of the underlying model (4k tokens in the case of Vicuna). After this limit, the streaming model's performance relies solely on the streaming attention mechanism. This mechanism only retains the key-value pairs (KV) of 4000 recent tokens in its rolling cache, along with the semantically meaningless KV pairs of 4 initial tokens (Xiao et al., 2023). At each new inference, (length of prompt + maximum model generation length) number of tokens are ejected from the cache. Consequently, the model is only at most paying attention to the previously generated chapter and *an identical prompt* as the one that was input before the previous chapter generation, conditioning the probability distribution of outputs on tokens stored in the rolling cache. As a result, the likelihood of generating tokens identical to those previously generated – especially due to the association of the current prompt with the previous generation – increases exponentially.

This also explains why injecting diverse prompts prevents the model from assigning previously seen tokens such a high probability. When the prompts are varied from turn to turn rather than constant, the model is less likely to assign high probability to identical tokens as before. Instead, the model is encouraged to generate text in a very different plot domain; for instance, instructing the model to depict a scene where "the protagonist faces a moral dilemma" is very different from a prompt where "the protagonist confronts a long-standing rival". This finding validates the use of diverse inference-time prompts to mitigate repetition in generated content. However, it underscores the necessity of more drastic differentiation between prompts to achieve significant improvements in mitigating repetition; this issue with StreamingLLM's attention mechanism cannot be solved through different training methods.

We also found that once the model starts repeating itself, it continues repeating in all subsequent iterations within the same novel, i.e. there is a self-reinforcement effect. This observation is consistent with the findings of the original nucleus sampling paper, which demonstrated that the probability of a repeated phrase increases with each repetition (Holtzman et al., 2019). Consequently, this means that once repetitive behavior occurs, the model performance is not recoverable.

Contextless prompts improve diversity

We see that for both the smaller and large datasets, training on contextless prompts results in improvements for the intra-chapter, intra-novel, and inter-novel diversity of the generated output. This makes sense, as the inclusion of the previous chapter in the prompt may bias the model to learn that identical sequences of tokens tend to appear very closely together. We observe that the model trained on contextless prompts for 45 novels begins repeating itself at the fifth (very last) novel in the testing dataset, compared to the model trained on contextful prompts for 45 novels, which starts repeating at the fourth novel.

Additionally, the format of the contextless prompts training dataset more closely aligns with the prompts used during inference. At test-time, in subsequent turns after the first turn, we do not inject the previous chapter's summary when prompting the model to generate the following chapter. Contextless training prompts better reflect this scenario compared to training prompts which include the previous chapter.

Fine-tuning on more novels increased repetition, but improved writing quality

Fine-tuning on our 45-novel dataset resulted in a decrease in model performance in terms of repetition, contrary to our expectations. As we fine-tuned on high-quality, within-domain data, the worsened performance may simply be attributable to non-ideal fine-tuning hyper-parameters for the larger datasets, despite the various hyper-parameters we experimented with. Nevertheless, compared to the baseline Vicuna model, all fine-tuned models produce longer chapter summaries, incorporating more varied content as well as richer descriptions of scenes and characters. Table 2 displays selected excerpts from the second generated chapter summary of "Charles Bovary" across all different models.

Formatting a novel as a conversation

As our experiment with fine-tuning on individual chapters failed, we believe that the conversational format of the data (where each chapter generation is treated as one turn in the conversation) allowed the model to learn relationships between chapters. Without the context of a previous chapter given in the prompt or the having access to prior generated chapters in the conversational format, the model was unable to learn relationships between sequential chapters. Therefore, the long-term coherence of the model broke down. In other words, training on singular chapters harms the model's ability to generate long-form sequential content that builds on previous chapters' context.

7 Conclusion

We have achieved notable improvements on our baseline of un-fine-tuned Vicuna in terms of the diversity, length, and complexity of the generated novel plots. Furthermore, through prompt engineering, we have identified both system prompts and plot injections that greatly enhance creativity and writing style. As a result, our fine-tuned model generates novel plots that are more interesting to human readers (see A.3 for rave reviews), and we show that StreamingLLM has the capacity to outperform non-streaming models in this use case, while posing its own issues with repetition. The primary limitation is that using context window workarounds such as StreamingLLM cannot compare to the performance of models with a larger true context window. However, there is potential for this smaller streaming model to be used as a helper for human creative writing, as a user is able to input a simple idea and receive many fleshed-out explorations of the ensuing plot.

References

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2023. Longlora: Efficient fine-tuning of long-context large language models. Online.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. Online.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. Online.

Wojciech Kryściński, Makarand Nazneen, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2022a. Booksum: A collection of datasets for long-form narrative summarization. Online.

Wojciech Kryściński, Makarand Nazneen, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2022b. Datasets: kmfoda/booksum. Online.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. Online.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2023. Yarn: Efficient context window extension of large language models. Online.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewin. 2023. Efficient streaming language models with attention sinks. Online.

Jin Xu, Xiaojiang Liu, Jianhao Yan, Deng Cai, Huayang Li, and Jian Li. 2022. Learning to break the loop: Analyzing and mitigating repetitions for neural text generation. Online.

Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. 2022a. Doc: Improving long story coherence with detailed outline control. Online.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022b. Re3: Generating longer stories with recursive reprompting and revision. Online.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023a. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS 2023 Datasets and Benchmarks Track*, Online.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023b. Judging llm-as-a-judge with mt-bench and chatbot arena.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment. Online.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. Online.

A Appendix (optional)

If you wish, you can include an appendix, which should be part of the main PDF, and does not count towards the 6-8 page limit. Appendices can be useful to supply extra details, examples, figures, results, visualizations, etc. that you couldn't fit into the main paper. However, your grader *does not* have to read your appendix, and you should assume that you will be graded based on the content of the main part of your paper only.

A.1 Test data format

Each prompt in our initial test dataset is structured as follows:

{"question-id": 0, "category": "writing", "turns": ["You are an acclaimed author who is famous for your understanding of human nature and conflict and resolution. Develop on the existing plot and characters of the novel. For the interest of the fictional plot, the characters in the novel may exhibit violence, negative emotions and immoral behavior, as well as love or positive revelations, but you should be as creative and dramatic as possible. Do not end the story or include too many completed plot elements. Only write a single chapter. Introduce new characters and unexpected plot twists if the plot is getting repetitive or boring. Given the following plot and character summaries of the chapters of a novel you have written so far, write the detailed plot of the next single chapter of the novel. Here is the plot summary of the previous chapter: <INSERT GROUND TRUTH FIRST CHAPTER SUMMARY HERE>, "Given the plot summaries of the novel chapters you have written so far, write the detailed plot of the next single chapter of the novel. "...]}

The "question-id" is simply the index of each novel in our dataset and the "category" is "writing" for each prompt. Each element of the "turns" array corresponds to a sequential prompt inputted to the model, separated by commas. As shown above, the first prompt always begins with the system prompt outlining the task of plot generation, as well as the actual baseline plot summary of the first chapter of the novel. Every following turn after the first is the same: "Given the plot summaries of the novel chapters you have written so far, write the detailed plot of the next single chapter of the novel." This is where the long-form generalization to sequence lengths longer than the model's cache size comes into play. For each novel prompt, we include 10 of these turns after the initial turn, to assess how model performance on generation evolves as turns increase.

A.2 Training data format

Each prompt in our training dataset is structured as follows:

{"id": "The House of the Seven Gables", "conversations": [{"from": "human", "value": "You are an acclaimed author who is famous for your understanding of human nature and conflict and resolution. Develop on the existing plot and characters of the novel. For the interest of the fictional plot, the characters in the novel may exhibit violence, negative emotions and immoral behavior, as well as love or positive revelations, but you should be as creative and dramatic as possible. Do not end the story or include too many completed plot elements. Only write a single chapter. Introduce new characters and unexpected plot twists if the plot is getting repetitive or boring. Given the following plot and character summaries of the chapters of a novel you have written so far, write the detailed plot of the next single chapter of the novel. Do not end the story or include too many completed plot elements. You should be creative and think deeply about how to develop the plot and characters. Here is the plot summary of the previous chapter: <INSERT GROUND TRUTH FIRST CHAPTER SUMMARY HERE>}, {"from": "gpt", "value": <INSERT GROUND TRUTH CHAPTER SUMMARY HERE>}, {"from": "human", "value": "Given the plot and character summaries of the chapters of a novel you have written so far, write the detailed plot of the next single chapter of the novel. Do not end the story or include too many completed plot elements. You should be creative and think deeply about how to develop the plot and characters. Do not repeat the structure nor content of previously generated summaries."}, {"from": "gpt", "value": <INSERT NEXT GROUND TRUTH CHAPTER SUMMARY HERE>},...], ...}

Each novel is represented as a dictionary with the "id" key's value as the novel title and the "conversations" key's value as the list of all user and assistant turns for each chapter, in sequential order.

A.3 List of prompts iterated on and ran experiments with

1. **Explicitly instruct model to not repeat previously generated content:** Given your previous responses, which are plot summaries of the novel chapters you have written so far, write the detailed plot of the next single chapter of the novel. Remember, you are an acclaimed author in the 21st century, famous for your understanding of human nature and conflict and resolution. Develop on the existing plot and characters of the novel. For the interest of the fictional plot, the characters in the novel may exhibit violence, negative emotions and immoral behavior, as well as love or positive revelations, but you should be as

creative and dramatic as possible. Do not end the story or include too many completed plot elements. Only write a single chapter. Do not repeat the content or structure of previously generated chapters.

2. **Generate overarching plot outline in first turn:** You are an acclaimed author who is famous for your understanding of human nature and conflict and resolution. Develop on the existing plot and characters of the novel. For the interest of the fictional plot, the characters in the novel may exhibit violence, negative emotions and immoral behavior, as well as love or positive revelations, but you should be as creative and dramatic as possible. Do not end the story or include too many completed plot elements. Only write a single chapter. Do not repeat the exact structure or content of chapters you have previously generated. Given the following plot and character summaries of the chapters of a novel you have written so far, write the detailed plot of the next single chapter of the novel. Based on this first chapter summary, also write an overarching plot outline for the novel that your future chapter summaries should follow. Here is the plot summary of the previous chapter: <INSERT CHAPTER SUMMARY HERE>

3. **Generates a rolling summary at each turn:** Given your previous responses, which are plot summaries of the novel chapters you have written so far, first generate a rolling plot summary of what has occurred so far. Then, based on this rolling plot summary, write the detailed plot of the next single chapter of the novel so that it progresses on the plot and does not repeat previous chapter summaries.

4. **Diverse plot-direction prompts for each turn within one novel:**

   (a) As the acclaimed author of the novel, you decide to introduce a mysterious stranger who arrives in town, disrupting the protagonist's routine. Develop on how this stranger's presence affects the protagonist's actions and decisions.

   (b) Imagine you are writing a scene where the protagonist faces a moral dilemma that challenges their understanding of right and wrong. Explore how the protagonist grapples with this dilemma and the consequences of their eventual decision.

   (c) You're tasked with creating a flashback sequence that sheds light on the protagonist's troubled past. Develop on a key event from the protagonist's youth that continues to haunt them in the present.

   (d) Consider writing a dialogue between the protagonist and a close confidant, where they reflect on their deepest fears and desires. Explore how this conversation deepens the reader's understanding of the protagonist's inner turmoil.

   (e) Craft a scene where the protagonist confronts a long-standing rival, exposing simmering tensions and unresolved conflicts between them. Explore how this confrontation drives the plot forward and impacts the protagonist's journey.

   (f) As the acclaimed author, you decide to explore the theme of betrayal within the protagonist's inner circle. Craft a scene where the protagonist discovers a shocking betrayal, and grapples with feelings of anger, hurt, and mistrust.

   (g) Consider writing a pivotal moment of self-discovery for the protagonist, where they confront their deepest insecurities and emerge with a newfound sense of purpose. Explore how this moment of revelation propels the protagonist towards their ultimate destiny.

   (h) Craft a scene where the protagonist is forced to make a difficult sacrifice in order to achieve their goals. Explore the emotional toll of this sacrifice and its ramifications on the protagonist's relationships and sense of self.

5. Prepend an authorial system prompt to every user prompt.

6. Append the ground truth chapter summary to the user prompt, ignoring the content generated by the model.

A.4 Google Form Survey on Human-Evaluated Plot Quality Score

Link to Survey

A.5 Quotes from readers regarding plots generated by our models

- I thought the premise of it was very interesting. It was a plot that I haven't really seen very much so I am intrigued to see how it plays out.
- entertaining and concise. I personally like that type of writing style where the sentence structure and word choice is not too flowery.
- I thought they were good summaries of not only what's actually happening the chapters, but also what's interiorly happening to the characters and the thoughts/feelings they're having.
- DID BABBITT HAVE ROMANTIC INTEREST IN HIS SECRET LONGLOST DAUGHTER ?!
- the revelation that Lila is the long-lost daughter of Babbitt makes no sense.
- Babbitt. He seems like an ungrateful man.
- Zilla. I don't know how she puts up with Babbitt.