# CHAPTERBREAK: A Challenge Dataset for Long-Range Language Models

**Simeng Sun** **Katherine Thai** **Mohit Iyyer**
University of Massachusetts Amherst
{simengsun,kbthai,miyyer}@cs.umass.edu

## Abstract

While numerous architectures for *long-range language models* (LRLMs) have recently been proposed, a meaningful evaluation of their discourse-level language understanding capabilities has not yet followed. To this end, we introduce CHAPTERBREAK, a challenge dataset that provides an LRLM with a long segment from a narrative that ends at a *chapter boundary* and asks it to distinguish the beginning of the ground-truth next chapter from a set of negative segments sampled from the same narrative. A fine-grained human annotation reveals that our dataset contains many complex types of chapter transitions (e.g., parallel narratives, cliffhanger endings) that require processing global context to comprehend. Experiments on CHAPTERBREAK show that existing LRLMs fail to effectively leverage long-range context, substantially underperforming a segment-level model trained directly for this task. We publicly release our CHAPTERBREAK dataset to spur more principled future research into LRLMs.[1]

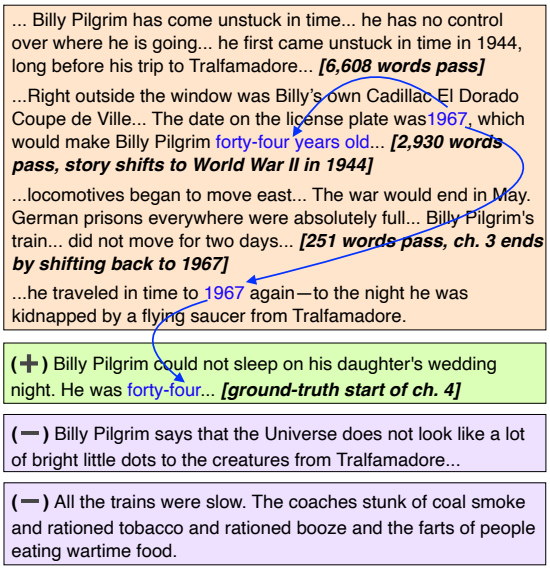


Figure 1: An illustrative example of our suffix identification task from Kurt Vonnegut's *Slaughterhouse-Five*, in which an LRLM needs to make connective inferences across temporal and spatial shifts in a long prefix of the narrative to correctly disambiguate the (+) start of the next chapter from (-) negative examples.

## 1 Introduction

Research on *long-range language models* (LRLMs) aims to process extremely long input sequences by making the base Transformer architecture more efficient (e.g., through sparse attention, recurrence, or cached memory). These modifications are commonly validated by training LRLMs on PG-19 (Rae et al., 2020), a long-document language modeling dataset, and demonstrating small perplexity decreases over shorter context models (Roy et al., 2021; **?**). However, recent analysis experiments (Sun et al., 2021; Press et al., 2021) show that modern LRLMs rely mostly on local context (i.e., the immediately preceding 1-2K tokens) and are insensitive to various perturbations applied to more distant context.

In this paper, we move beyond token-level perplexity by evaluating LRLMs on a task that requires a rich understanding of long-range dependencies. Our task is an instance of *suffix identification*, in which a language model is given a long input sequence (or *prefix*) and asked to disambiguate the next $n$-token segment from a set of hard negatives sampled from the same narrative. To succeed at this task, an LRLM should assign high probability to the ground-truth next segment and low probability to the negatives. To specifically test long-range dependencies, we restrict our prefixes to end at *chapter breaks* of a longer cohesive narrative (e.g., a novel).

We construct a challenge dataset, CHAPTERBREAK, by automatically detecting chapter bound-

[1]We make our code and data public at https://github.com/SimengSun/ChapterBreak



*Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3704 - 3714

aries within both held-out PG-19 documents (in-domain for pretrained LRLMs) and works of fan fiction published on the Archive of Our Own (out of domain).[2] We perform a detailed analysis of the types of chapter transitions in our dataset and discover a high frequency of narrative shifts in point-of-view, location, and time, all of which require global narrative understanding over long input sequences. For example, Figure 1 contains a complex prefix in which the time-traveling Billy Pilgrim moves between World War II, 1960s suburban life, and an alien planet. Understanding the cliffhanger ending, in which the narrative abruptly switches from a wartime scene to a 1967 alien abduction, requires an LRLM to make connective inferences using details buried far back in the context (e.g., Billy's age in 1967).

We evaluate three LRLMs on CHAPTERBREAK, including BigBird (Zaheer et al., 2020), the Routing Transformer (Roy et al., 2021), and its local attention variant, all pretrained or fine-tuned on PG-19. Our experiments show that these LRLMs perform poorly at selecting the ground-truth suffix, regardless of the length of the input sequence. As an upper bound, we train a small RoBERTa-based segment-level language model on PG-19 and discover that it substantially outperforms all LRLMs on CHAPTERBREAK, which suggests that LRLMs have considerable room for improvement on this suffix identification task. Finally, we perform an analysis on the instances for which all models struggle to choose the correct suffix, which shows that shifts in location and events in focus are particularly challenging to disambiguate. Taken together, these results suggest that CHAPTERBREAK is a useful benchmark for future research into LRLMs.

## 2 The CHAPTERBREAK dataset

Authors often break long-form narratives into a sequence of discrete chapters to impose "an order and shape over events in time" (Stevick, 1970). Henry Fielding writes in his novel *Joseph Andrews* that the space between chapters is like "an Inn or Resting Place" for readers to reflect on the preceding chapter (Fielding, 1779). Chapters come in many flavors: for example, Murakami's *Kafka on the Shore* uses chapter breaks to alternate between parallel narratives focusing on the two protagonists, while cliffhanger endings such as the one in Figure 1 add suspense. Making sense of the complex narrative shifts associated with chapter transitions (e.g., changes in point-of-view, time, location, and theme) requires a deep understanding of the entire text. To maintain global narrative coherence, Myers et al. (1994) show that human readers tend to reactivate memory about "backgrounded" information from the long-range context.

**Task overview:** Given that chapter transitions requires global context understanding, how can we turn this into a task to evaluate LRLMs? A simple approach is to evaluate the token-level perplexity of an LRLM only at chapter boundaries (i.e., on the first $n$ tokens of each chapter); however, the vast majority of tokens can be predicted using just local context (Sun et al., 2021) under the teacher-forcing setup, which obscures an LRLM's usage of long-range context as we show in Section 3. We instead turn to the task of *suffix identification*, which closely resembles existing datasets such as SWAG (Zellers et al., 2018).

Each instance of our task is defined by a triplet $(c, s^+, s_i^- \in \mathbf{N})$, where $c$ is a prefix sequence of up to 8K tokens that ends at a chapter break, $s^+$ is the gold suffix of length 128 tokens (i.e., the beginning of the next chapter), and $s_i^-$ is a negative 128-token-long suffix from a set $\mathbf{N}$ of five[3] future chapter beginnings sampled from the same narrative.[4] All negatives are modified to begin with the same chapter index (e.g., if the gold suffix begins with "Chapter III", the chapter indices of all negatives is set to "Chapter III") to eliminate the effect found by Sun et al. (2021) of language models memorizing chapter indices in long contexts. We then evaluate whether an LRLM assigns higher probability to the gold suffix $P(s^+|c)$ than to all negative suffixes $P(s_i^-|c)$.

**Dataset overview:** Where do we get these triplets from? We collect a dataset, CHAPTERBREAK, with two splits: CHAPTERBREAK$_{PG19}$, which contains 241 examples extracted from the PG-19 validation set (Rae et al., 2020),[5] and CHAPTERBREAK$_{AO3}$, which contains 7,355 ex-

[2] https://archiveofourown.org

[3] We use a small number of negatives because it is time-consuming and resource-intensive to evaluate the probabilities of long sequences with LRLMs.

[4] In Appendix F, we show that in-book negatives are much harder than out-of-book negatives as they often contain the same named entities and rare tokens as the gold suffix. Thus, disambiguating the correct suffix requires a deep understanding of the context.

[5] We only collect examples from validation set as two baseline models in the later sections are trained on PG-19.

| Category | Definition | Pct. |
|---|---|---|
| Events | Previous event ends and new event starts | 76% |
| | Previous event continues into next chapter | 24% |
| Actors | Change of perspective or character in focus | 36% |
| | No change in POV or main character | 64% |
| Locations | Change of location | 68% |
| | No change in location | 32% |
| Continuity | Discontinuous but chronological | 29% |
| | Continuous | 62% |
| | Analepsis | 2% |
| | Parallel | 6% |

Table 1: Our human annotation on 300 chapter transitions randomly sampled from CHAPTERBREAK$_{AO3}$ shows the diversity and complexity of the dataset.

amples extracted from an online dump[6] of fanfiction posted on Archive of Our Own (AO3). We apply filtering to remove fanfiction works that are too short or not rated for general audiences. Each work contains on average 42K words and 21.5 chapters.[7] Even though the CHAPTERBREAK$_{PG19}$ split is small, we include it because many LRLMs are pretrained on PG-19; the much larger CHAPTERBREAK$_{AO3}$ split is out-of-distribution for all models that we evaluate. To extract chapters in PG-19, we match for lines beginning with the string "chapter", while AO3 stories already have chapter-level metadata.

**What are the different types of transitions in CHAPTERBREAK and how often do they occur?** To get a better sense of our dataset, we perform a fine-grained annotation of 300 randomly-selected chapter transitions from CHAPTERBREAK$_{AO3}$. For each transition, we annotate any changes in the following four aspects: events, actors (characters in focus), locations, and continuity. To annotate continuity, we follow a simplified version of the scheme proposed by Ireland (1986),[8] which considers five categories: **continuous** (the next chapter occurs within a day of the previous chapter), **discontinuous** (the next chapter occurs more than a day after the previous chapter), **analepsis** (the next chapter is a "flashback" to an earlier point in the narrative), and **parallel** (the next chapter reverts to the time of a previous chapter, switching the character or event in focus).[9] The results, shown in Table 1, demonstrate that CHAPTERBREAK covers a diverse array of transitions, including many that require global narrative understanding.

[6] https://archive.org/download/AO3_story_dump_continuing

[7] More preprocessing details and statistics can be found in Appendix A.

[8] To validate our continuity annotations, we also annotate every chapter in *Pride and Prejudice* and obtain almost the same proportion of continuous transitions (67%) as the number reported by the expert annotation of Ireland (1986) (72%).

| | #Params | Seq Len | PPL$_{PG19}$ | Acc$_{PG19}$ | Acc$_{AO3}$ |
|---|---|---|---|---|---|
| LT | 516M | 8K | 76.8 | 25% | 24% |
| RT | 490M | 8K | 72.3 | 22% | 24% |
| Bigbird | 128M | 4K | 56.2 | 27% | 26% |
| GPT-2 | 1.5B | 1K | 78.2 | 23% | 24% |
| GPT-3 | 175B | 2K | - | 36%* | 28%* |
| SuffixLM | 87M | 10K | - | **52%** | **41%** |

Table 2: Summary of LRLMs (top), Transformer LMs (middle), and our SuffixLM (bottom). All models are trained or fine-tuned on PG-19 except for GPT-2. The third column shows the word-level perplexity of gold suffix in the PG-19 split. The last two columns show the suffix identification accuracy of each model on the two CHAPTERBREAK splits when evaluated at maximum input length. * indicates results are on a subset of CHAPTERBREAK.

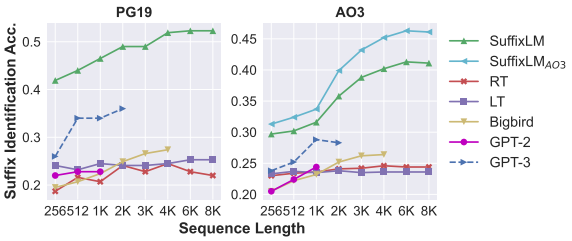


Figure 2: Suffix identification accuracy on both splits (PG-19 and AO3) of CHAPTERBREAK is much lower for LRLMs than our SuffixLM upper bound.

## 3 Experiments

We evaluate three different long-range language models on CHAPTERBREAK and compare their results to those of standard Transformer language models as well as an upper bound directly trained for suffix prediction.

**Language models:** We evaluate three LRLMs pretrained on PG-19: the Local Transformer (Roy et al., 2021, LT), Routing Transformer (RT) (Roy et al., 2021, RT), and BigBird (Zaheer et al., 2020). The BigBird model is the decoder part of the released checkpoint fine-tuned with causal LM objective on 14k books of PG-19 for 100k steps. We also evaluate two standard Transformer language models, GPT-2 large (Radford et al., 2019) and GPT-3 (Brown et al., 2020).[10] We summarize these

[9] Appendix B contains more details about each category.

[10] Due to OpenAI's API costs for GPT-3, we only evaluate in total a subset of 200 examples instead of the full dataset.

models in Table 2, more details about each model are included in Appendix C.

**An upper bound directly trained for suffix identification:** As authors often write stories that are intended to surprise readers, it is possible that many examples in CHAPTERBREAK are ambiguous by nature (i.e., the upper bound for suffix identification accuracy may not be 100%). To obtain a reasonable upper bound, we also train a model (SuffixLM) directly on the suffix identification task by scaling up the sentence-level language model proposed by Ippolito et al. (2020).[11] We divide an input sequence into multiple segments, each of which is embedded via the `[CLS]` vector of a small fine-tuned RoBERTa network (Liu et al., 2019). Our SuffixLM then performs "language modeling" atop the dense `[CLS]` vectors, predicting the next segment representation given the representations of previous segments via contrastive predictive coding (van den Oord et al., 2018).[12] Formally, our SuffixLM minimizes the following loss:

$$\mathcal{L}_i = -\log \frac{\exp(\hat{\mathbf{z}}_i{}^\top \mathbf{z}_i^+)}{\sum_{\mathbf{z}_i \in \{\mathbf{z}_i^+, \mathcal{Z}_i^-\}} \exp(\hat{\mathbf{z}}_i{}^\top \mathbf{z}_i)}$$

where $\hat{\mathbf{z}}_i$ is the predicted representation by SuffixLM, $\mathbf{z}_i^+$ is the gold suffix representation obtained from a small encoder (RoBERTa), and $\mathcal{Z}_i^-$ is the set of dense representations of the negatives. More details about our SuffixLM are included in Appendix D.

## 4 Results & Analysis

Overall, the results in Table 2 (rightmost two columns) confirm that all of the language models studied in this paper struggle on CHAPTERBREAK, especially when compared to the SuffixLM upper bound, which outperforms the best LM by ∼25% absolute accuracy when evaluated on the entire PG-19 split. We describe other interesting results and analysis below:

**Accuracy increases with longer prefixes:** Figure 2 shows that as prefix sequence length increases, some LRLMs (e.g., LT) barely improve, while others show modest improvements (e.g., GPT-3 and fine-tuned BigBird). However, all LRLMs significantly underperform our SuffixLM upper bound, even when the SuffixLM is given prefixes that are only 256 tokens long. Additionally, SuffixLM's accuracy increases far more than those of LRLMs when increasing the prefix length (from 31% at prefix length of 256 to 46% at 8K on the AO3 split[13]). This result suggests that the token-level LRLMs evaluated in our work are not taking full advantage of information in the long-range context to solve CHAPTERBREAK.

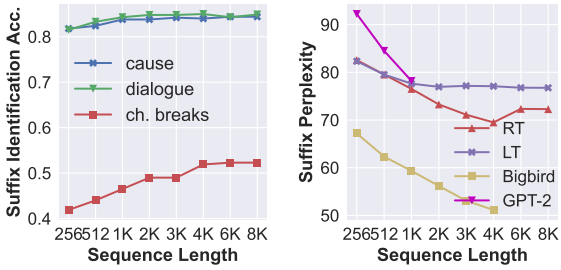


Figure 3: **Left:** Prefixes ending at chapter breaks benefit more from long-range context than other types of discourse boundaries. **Right:** Word-level perplexity of the gold suffix does not correlate to accuracy (e.g., GPT-2 has high perplexity but outperforms RT on suffix identification).

**Perplexity does not always correlate with accuracy:** Previous LRLM efforts use validation perplexity (e.g., on PG-19) to compare against other models. However, we show that perplexity is not by itself a predictor of suffix identification accuracy: As shown in Table 2, GPT-2 achieves higher accuracy than RT despite yielding a word-level perplexity of 78.2 on gold suffixes, compared to 72.3 for RT.[14] We advocate that future research on LRLMs includes evaluation on suffix identification tasks like CHAPTERBREAK, as perplexity alone does not reflect LRLMs' capabilities to model long-range dependencies.

**Why chapter breaks over other discourse boundaries?** Other discourse markers, including *cause* and *dialogue*, also often prompt human readers to reactivate memories of global context (Albrecht

[11] Our SuffixLM can process up to 10K tokens, while the model of Ippolito et al. (2020) supports only up to ten sentences.

[12] Our SuffixLM is closely related to the model in Ainslie et al. (2020), but differs crucially by predicting the representation of next segment instead of summaries.

[13] We collected 13,682 fan-fictions posted on AO3 and fine-tuned our SuffixLM on subset of this dataset to be the model SuffixLM$_{AO3}$. More details about the filtered AO3 works are included in Appendix A

[14] As these models use different tokenizers, we normalize the subword-level perplexities to the word level as suggested by Rae et al. (2020). More details about this can be found in Appendix E.

and Myers, 1995). We create suffix identification datasets for these two discourse markers by string matching over corresponding cue phrases ('because', 'due to' for the *cause* subset and text within quotation marks for *dialogue*).[15] Figure 3 (left) shows that with prefixes of length 256 tokens, our SuffixLM is able to successfully disambiguate the correct suffixes for both discourse markers more than 80% of the time, while the accuracy is much lower at chapter boundaries. As the prefix length increases, accuracy only slightly increases for *cause* and *dialogue*, especially compared to the robust improvement at chapter boundaries.[16]

**Short-context Transformers are comparable to LRLMs:** Our results show that GPT-2, despite its high perplexity on gold suffixes and short maximum sequence length (1024 tokens), achieves comparable performance to RT and LT on both splits. Meanwhile, GPT-3 achieves much higher performance on both CHAPTERBREAK at a sequence length of 2,048 tokens, and the increasing GPT-3 curve in Figure 2 is promising for future work scaling LMs to longer sequence lengths.

**Limitations of our work:** While we have used the SuffixLM as an upper bound in this paper and demonstrated that it substantially outperforms LRLMs on CHAPTERBREAK, a more compelling comparison would include human performance on our task at varying prefix lengths, especially since some chapter transitions are specifically intended by their authors to be unpredictable. However, obtaining reliable human performance numbers is very difficult, as it requires in-depth comprehension of long narratives on the part of workers. Due to the time-consuming nature of this task and its high cognitive demand, it is not possible (within a reasonable budget) to use crowdsourcing, as ensuring that the annotators fully read the prefix instead of skimming or ignoring it is a major challenge. These issues also carry over to experiments performed with in-person subjects. As such, we leave a thorough human evaluation on CHAPTERBREAK to future work.

## 5 Related Work

Our work depends heavily on recent advances in efficient Transformers (Tay et al., 2020) that process long sequences (Rae et al., 2020; Beltagy et al., 2020; Zaheer et al., 2020; Ainslie et al., 2020; Roy et al., 2021). Sparse attention (Child et al., 2019), relative position encoding (Shaw et al., 2018; Raffel et al., 2020; Guo et al., 2021), recurrence mechanism and memory (Dai et al., 2019; Weston et al., 2015; Hutchins et al., 2022; **?**) and other tricks (Shen et al., 2020; Katharopoulos et al., 2020; Gupta and Berant, 2020; Stock et al., 2021; Yogatama et al., 2021; Borgeaud et al., 2021; Hawthorne et al., 2022) are commonly adopted by recent Transformer variants to make the operation on long sequences more time/memory efficient.

Besides perplexity, many downstream extrinsic tasks for evaluating long-range language models were developed recently , such as long-form QA (Fan et al., 2019; Pang et al., 2021), document-level summarization (Kryściński et al., 2021; Huang et al., 2021), and machine translation (Liu and Zhang, 2020). More recently, Shaham et al. (2022) introduce a new benchmark covering multiple domains and tasks, while Tay et al. (2021) propose multimodal long sequence tasks.

## 6 Conclusion

We introduce CHAPTERBREAK, a suffix identification dataset targeted at evaluating the discourse-level understanding of long-range language models. The dataset is extracted from long-form narratives and covers a variety of complex chapter transitions, such as shifts in location and events in focus. Experiments show that existing LRLMs perform poorly on CHAPTERBREAK and much worse than a SuffixLM trained as an upper bound on this task. We release the dataset to spur more principled development of future LRLMs.


## Acknowledgements

We thank the anonymous reviewers and UMass NLP group for the thoughtful comments on the draft of this paper. We are grateful to AO3 Support Chair and volunteers for answering data related questions. This work was supported by awards IIS-1955567 and IIS-2046248 from the National Science Foundation (NSF).


## Ethical Considerations

CHAPTERBREAK is constructed from two sources: public domain books published prior to 1919 (from the held-out set of PG-19) and works of fanfiction extracted from an online dump of stories posted

[15] Appendix A contains more details about data for these two discourse markers.

[16] Appendix G shows similar trends on *cause* and *dialogue* with other models.

on Archive of Our Own (AO3). We refer readers to Rae et al. (2020) for more details about PG-19. For AO3, we apply multiple filters to obtain long fanfiction stories rated as suitable for "General Audiences". We refer readers to Appendix A for more preprocessing details. More generally, this work focuses on long-range language models, which could potentially be misused to generate offensive output. However, the main purpose of this paper is to present a dataset which provides a better evaluation of the discourse-level capabilities of such models.

## References


Joshua Ainslie, Santiago Ontanon, Chris Alberti, Vaclav Cvicek, Zachary Fisher, Philip Pham, Anirudh Ravula, Sumit Sanghai, Qifan Wang, and Li Yang. 2020. Etc: Encoding long and structured inputs in transformers.

Jason E. Albrecht and Jerome L. Myers. 1995. Role of context in accessing distant information during reading. *Journal of experimental psychology. Learning, memory, and cognition*, 21 6:1459–68.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2021. Improving language models by retrieving from trillions of tokens. *arXiv preprint arXiv:2112.04426*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering.

Henry Fielding. 1779. *The History of the Adventures of Joseph Andrews..*, volume 1. J. Fr. Valade.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2021. Longt5: Efficient text-to-text transformer for long sequences.

Ankit Gupta and Jonathan Berant. 2020. Gmat: Global memory augmentation for transformers. *arXiv preprint arXiv:2006.03274*.

Curtis Hawthorne, Andrew Jaegle, Cătălina Cangea, Sebastian Borgeaud, Charlie Nash, Mateusz Malinowski, Sander Dieleman, Oriol Vinyals, Matthew Botvinick, Ian Simon, et al. 2022. General-purpose, long-context autoregressive modeling with perceiver ar. *arXiv preprint arXiv:2202.07765*.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, Online. Association for Computational Linguistics.

DeLesley S. Hutchins, Imanol Schlag, Yuhuai Wu, Ethan Dyer, and Behnam Neyshabur. 2022. Block-recurrent transformers. *ArXiv*, abs/2203.07852.

Daphne Ippolito, David Grangier, Douglas Eck, and Chris Callison-Burch. 2020. Toward better storylines with sentence-level language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7472–7478, Online. Association for Computational Linguistics.

KR Ireland. 1986. Towards a grammar of narrative sequence: The model of the french lieutenant's woman. *Poetics today*, 7(3):397–420.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and Francois Fleuret. 2020. Transformers are rnns: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning*.

Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2021. Booksum: A collection of datasets for long-form narrative summarization.

Siyou Liu and Xiaojun Zhang. 2020. Corpora for document-level neural machine translation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3775–3781, Marseille, France. European Language Resources Association.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019.

Roberta: A robustly optimized bert pretraining approach.

Jerome Myers, Edward O'Brien, Jason Albrecht, and Robert Mason. 1994. Maintaining global coherence during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20:876–886.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, and Samuel R. Bowman. 2021. Quality: Question answering with long input texts, yes!

Ofir Press, Noah A. Smith, and Mike Lewis. 2021. Shortformer: Better language modeling using shorter inputs. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5493–5505, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. 2020. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. 2021. Efficient content-based sparse attention with routing transformers. *Transactions of the Association for Computational Linguistics*, 9:53–68.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, and Omer Levy. 2022. Scrolls: Standardized comparison over long language sequences.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, New Orleans, Louisiana. Association for Computational Linguistics.

Sheng Shen, Zhen Dong, Jiayu Ye, Linjian Ma, Zhewei Yao, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. 2020. Q-bert: Hessian based ultra low precision quantization of bert. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8815–8821.

Philip Stevick. 1970. *The Chapter in Fiction: Theories of Narrative Division*. Syracuse, NY: Syracuse University Press, c1970.

Pierre Stock, Angela Fan, Benjamin Graham, Edouard Grave, Rémi Gribonval, Herve Jegou, and Armand Joulin. 2021. Training with quantization noise for extreme model compression. In *International Conference on Learning Representations*.

Simeng Sun, Kalpesh Krishna, Andrew Mattarella-Micke, and Mohit Iyyer. 2021. Do long-range language models actually use long-range context? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 807–822, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2021. Long range arena : A benchmark for efficient transformers. In *International Conference on Learning Representations*.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2020. Efficient transformers: A survey.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2015. Memory networks.

Dani Yogatama, Cyprien de Masson d'Autume, and Lingpeng Kong. 2021. Adaptive semiparametric language models. *Transactions of the Association for Computational Linguistics*, 9:362–373.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

## A Dataset statistics

We collected 13,682 fanfictions from an online dump of stories posted on Archive of Our Own (AO3) by filtering works written in English language, rated General Audience by the author and contains at least 10K words and more than 10 chapters. For each chapter, we remove the text within the range of "`**Notes for the Chapter:**`", "`**Summary for the Chapter:**`" and "`**Author's Note:**`". The meta-comments inserted into the main text by the authors are not removed. The statistics of this long-fic dataset are included in Table 3. We do not apply other profanity filters to the fictions, therefore there may still be inappropriate content for general audience as the rating is self-labeled by each author. Besides chapter breaks introduced in the main text, we also collected two other discourse boundaries, cause and dialogue, as comparisons to the chapter boundary examples. We present the statistics each type of examples in Table 4.

- **Cause**: The beginning of the suffix contains words or phrases 'because', 'due to', 'owing to'. According to (Albrecht and Myers, 1995), human readers reactivate memory of global context for comprehending statements following causes or goals.

- **Dialogue**: The gold suffix in this category starts with a quotation mark. This often happens in dialogues where the continuation of one interlocutor depends heavily on the immediately preceding utterance. We conjecture this is the type where the prediction relies more on the local rather than the global context.

| | mean | min | max |
|---|---|---|---|
| #chapters | 21.5 | 11 | 589 |
| #words | 41,513.2 | 10,000 | 636,468 |

Table 3: Statistics of long fanfictions collected from AO3 story dump.

## B Annotation Scheme

We annotate each chapter transition from four aspects: events, actors (point-of-view or characters in focus), location, and continuity in timeline.

| | AO3 | | PG19 | |
|---|---|---|---|---|
| Suffix Type | #works | #examples | #works | #examples |
| cause | 965 | 8,133 | 45 | 506 |
| dialogue | 979 | 8,724 | 46 | 3,165 |
| chapter breaks | 1202 | 7,355 | 17 | 241 |

Table 4: Data statistics of CHAPTERBREAK as well as another two discourse boundary examples.

**Events** We define two subcategories based on whether (1) previous event ends in the previous chapter and new event starts in the new chapter, (2) old event does not end and continues into the next chapter.

**Actors** We define two subcategories based on whether there is a shift in POV or main character in focus.

**Location** We define two subcategories based on whether the location described in the prefix and in the new chapter is different.

**Continuity** Following Ireland (1986)'s work, we categorize the chapter transition by timeline continuity into four subcategories:

- **Discontinuous but chronological**: Reusing the standard by Ireland (1986), discontinuous represents a gap in time forward for more than one night.

- **Continuous**: The time interval between chapters lasts for no more than one night.

- **Analepsis**: Analepsis represents retrospective evocation of an event, or "flashback" to an earlier point in the narrative.

- **Parallel**: This includes timeline reverting back to the time of any previous chapter, typically accompanied by switching character in focus or description of a separate set of events independent of the last chapter. This category is a collapse of "alternate phase", "parallel phase" and "simultaneous phase" introduced in (Ireland, 1986).

## C Baselines

**Bigbird (Zaheer et al., 2020)** To reduce the quadratic complexity of self-attention in the standard Transformer, the Bigbird model employs a mixture of global, random and local attention mechanisms, which successfully reduce the complexity

to linear. The idea is to insert each sequence $O(1)$ global tokens, which attend to all other tokens. The rest tokens attend to their neighbor tokens, random tokens in the sequence as well as the inserted global tokens. A very similar idea is developed concurrently in the Longformer (Beltagy et al., 2020). The Bigbird model we fine-tuned is the decoder part of the released checkpoint. We fine-tune the model with causal LM objective on 14K books of PG-19 with peak learning rate 0.0001 for 100K steps. We set attention type to be "original_full" instead of using "block_sparse" during fine-tuning. Training is completed on a single RTX8000 GPU for around 6 days.

**Local Transformer** Rather than implementing all three types of sparse attention in Bigbird, the Local Transformer relies only on the local attention, i.e., each token attends to neighbors within a local window. The maximum attainable sequence length scales linearly with the number of layers, e.g., with window size $k$, the token representation at layer $l$ theoretically covers information in a range of $k \times l$ tokens.

**Routing Transformer (Roy et al., 2021)** Different from previously described models which use *position*-based sparse attention, the Routing Transformer employs *content*-based sparse attention. Namely, each token are routed to clusters and the attention is performed only within each cluster. The clustering operation effectively reduces the quadratic complexity in length $L$ to $O(L^{1.5})$. Both the RT and LT checkpoint we used were trained on PG-19 (Rae et al., 2020). For both RT and LT, we evaluate on single RTX8000 GPU.

**GPT-2/3** The GPT models have a lot shorter maximum input length than the rest models we evaluated. While GPT-2 model does not use sparse attentions at all, GPT-3 model adopts alternated layers of sparse and dense self-attention. We use the GPT-2 large model, which was pre-trained on data scraped from the Internet. The GPT-3 model was pre-trained on a mixture of filtered CommonCrawl, WebText2, Books1, Books2, and Wikipedia.

## D Finding the best SuffixLM

As there are no prior long-range segment-level LM architectures that we can borrow from, we experiment multiple design choices and report the result of only the best performing one in the main text. For all variants, we use RoBERTa-base (Liu et al., 2019) as the encoder to obtain the encoded segment representation. This is done by extracting the representation of the $[CLS]$ token prepended at the beginning of each sequence. We describe five variants below.

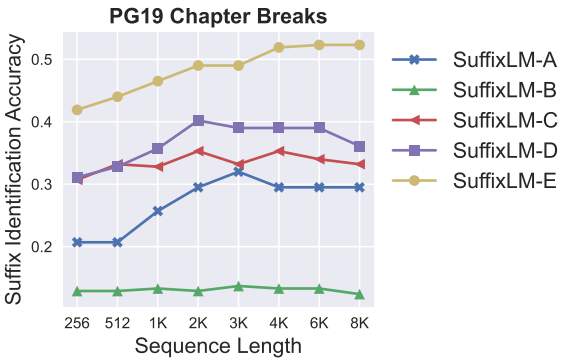


Figure 4: Performance of each SuffixLM variant. Detailed information about each variant is included in Appendix D.

- **SuffixLM-A** This variant contains a frozen RoBERTa-base encoder and a SuffixLM using a 6-layer Transformer as the base architecture.
- **SuffixLM-B** This variant contains a frozen RoBERTa-base encoder and a SuffixLM using a 6-layer average-attention Transformer as the backbone. The motivation of using uniform distribution for attention weights is to encourage the model to get more information from the distant context rather than rely too much on local context.
- **SuffixLM-C** This variant is essentically SuffixLM-A but during training we perform "segdrop" – stochastically dropping prefix segments with probability 0.2[17] when performing self-attention. When the local segments are dropped, the model has to predict the next segments with only the distant context, which also encourages learning better long-range prefix representations.
- **SuffixLM-D** Instead of freezing the encoder, this variant fine-tunes part of the encoder and the rest is the same as SuffixLM-A. Due to limited memory capacity, we only fine-tune the last two layers of the RoBERTa-base.
- **SuffixLM-E** This model is the same as SuffixLM-D except that we truncate the en-

[17] Tried {0.1, 0.2, 0.4}, 0.2 works the best.

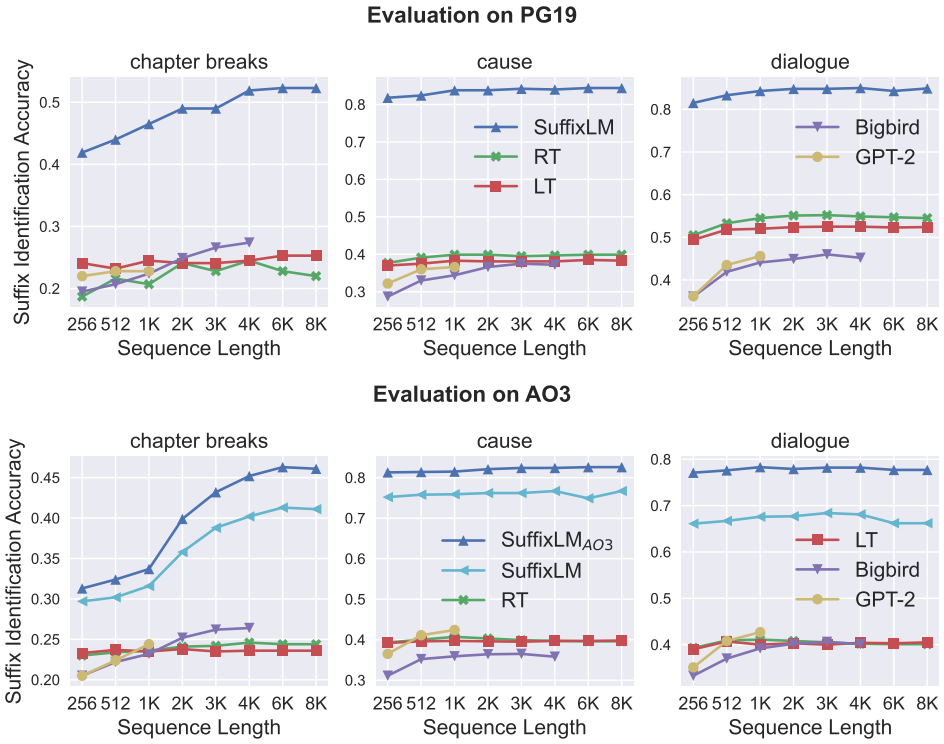


Figure 5: Evaluation results on both CHAPTERBREAK$_{PG19}$ and CHAPTERBREAK$_{AO3}$.

coder to just the two tunable layers and train all parameters in the encoder including the embedding parameters.

All SuffixLMs with frozen encoders are trained with average sequence length of 10240 tokens for up to 60k steps, and the one with trainable encoder is trained for max 120k steps. The dimension of the model is 768, hidden dimension 2048,attention heads 8. The peak learning rate is 0.0001 with warm up steps 4000. We train SuffixLM on entire PG-19 dataset and evaluate the best checkpoint selected by dev loss. We use segment size 128 in all SuffixLMs we trained. Each segment starts from a new sentence, if not reaching 128 tokens, we pad with a special '<pad>' token. For very long sentences, the part exceeding 128 tokens overflows to the next segment. We plot the suffix identification accuracy of each variant on CHAPTERBREAK while feeding in prefixes of increasing length. As shown in Figure 4, SuffixLM-E outperforms all other variants across various prefix lengths. Therefore in the main text, all SuffixLM refers to the SuffixLM-E variant. Note that one limitation of SuffixLM is it exclusively models on segment-level, which prohibits it from performing token-by-token generation and thus impossible for us to evaluate perplexity.

## E Suffix perplexity

Although the task of CHAPTERBREAK is to identify gold suffix from negatives, we also present the gold suffix perplexity of next-token prediction LMs. Note that all models were trained or fine-tuned on PG-19 except for GPT-2/3. As these models use different tokenizers, the 128-token suffix may cover different number of words, to make the results comparable, we convert the subword-level perplexity to word-level by multiplying a constant to the log probability value of each model. For RT/LT, we multiple by 1.248 as used in the official repository. We multiply the value by 1.30 for GPT-2, and 1.22 for Bigbird. These values are estimated via the subword/word ratio on validation set of PG-19. Our fine-tuned Bigbird model achieves the lowest perplexity on PG-19, even better than Routing Transformer or Local Transformer. This implies that context from long-range is not necessary for achieving low perplexity since the maximum input length of Bigbird is half that of RT/LT.

## F In-book vs. Out-of-book

This section is better read after reading through § 3. In this analysis experiment, we show why it is better that the negatives are from the same narrative as the gold suffix. We evaluate our upper bound

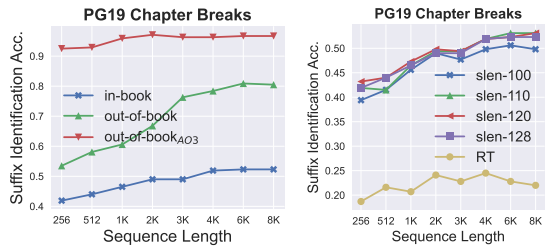


Figure 6: **Left:** In-book vs. out-of-book. **Right:** SuffixLM performance when evaluated with different suffix length. The variation in suffix length does not explain the large gap between SuffixLM and token-level LMs.

model SuffixLM on PG-19 set when the negatives are out-of-book suffixes, and plot the suffix identification accuracy in Figure 6. When evaluate against out-of-book negatives, this suffix identification task is almost solved by our SuffixLM, especially when the out-of-book examples are from another split in ChapterBreak. The extremely high accuracy under out-of-book setup suggests the segment representation from different books are easy for SuffixLM to distinguish, thus we adopt a harder setup where the negatives are from the same book. Besides, in-book negatives may contain the same re-occurring named entities or rare words, which require solid understanding of the prefix to differentiate the gold from the distractors.

## G Various Discourse Relationships

In addition to chapter breaks, we also evaluate the other two types of discourse boundary examples introduced in Appendix A. As shown in Figure 5, for all suffix types other than chapter breaks, the evaluated models stop improving as the sequence length grows to more than 2K tokens long. However, there is a significant increasing trend in chapter breaks for SuffixLM. For the rest models, the performance is either flat or not improving. On the AO3 split, the accuracy of SuffixLM improves for $\sim$ 15% as the sequence length increases from 256 to 8K, whereas the improvement of RT is only $\sim$ 1.4%. This is in contrast with SuffixLM's $\sim$ 1.5% and RT's $\sim$ 0.3% improvement for the 'cause' examples. We draw two conclusions from these observations: (1) the chapter breaks examples form a special case where longer prefix is preferred in order to pick the correct continuation. (2) By comparing the relative improvement, the token-level LMs fall far behind the SuffixLM, which is, besides the absolute performance gap, another evidence that current LRLMs do not effectively leverage long-range context for sequence tasks requiring discourse-level understanding.

## H Tackle difference in Tokenizers

As the models we evaluated use different tokenizers, there are small variations in term of suffix length, i.e., the 128-token suffix may cover different number of words. To understand how the difference in length impacts validity of evaluation, we evaluate SuffixLM with various suffix lengths. Figure 6 (right) indicates even though there are small variances when the suffixes are of different lengths, the large gap between SuffixLM and Routing Transformer still remains, thus the difference in suffix length does not explain the large performance gap.

## I Error analysis

**Models struggle with location and event shifts:** Among the 300 examples we annotated in Section 2, 89 examples were wrongly predicted by all models we have evaluated. By breaking the incorrectly predicted examples into category as presented in Table 1, we find that models tend to make wrong prediction when there is a shift in location or event, and when plots are continuous in timeline.[18]

| Category | Definition | Ratio |
|---|---|---|
| Events | Previous event ends and new event starts | 0.74 |
| | Previous event continues into next chapter | 0.26 |
| Actors | Change of perspective or character in focus | 0.43 |
| | No change in POV or main character | 0.57 |
| Locations | Change of location | 0.64 |
| | No change in location | 0.36 |
| Continuity | Discontinuous but chronological | 0.24 |
| | Continuous | 0.62 |
| | Analepsis | 0.03 |
| | Parallel | 0.11 |

Table 5: Human annotation on 89 examples sampled from ChapterBreak$_{AO3}$ where all models make the wrong prediction. 74% errors come from the examples where new event starts from the new chapter and 64% errors from the change of location.

[18] Detailed numbers are included in Appendix I.