

Do Long-Range Language Models Actually Use Long-Range Context?

Simeng Sun¹ Kalpesh Krishna¹ Andrew Mattarella-Micke² Mohit Iyyer¹

University of Massachusetts Amherst¹ Intuit AI²

{simengsun, kalpesh, miyyer}@cs.umass.edu

andrew_mattarella-micke@intuit.com

Abstract

Language models are generally trained on short, truncated input sequences, which limits their ability to use discourse-level information present in long-range context to improve their predictions. Recent efforts to improve the efficiency of self-attention have led to a proliferation of long-range Transformer language models, which can process much longer sequences than models of the past. However, the ways in which such models take advantage of the long-range context remain unclear. In this paper, we perform a fine-grained analysis of two long-range Transformer language models (including the *Routing Transformer*, which achieves state-of-the-art perplexity on the PG-19 long-sequence LM benchmark dataset) that accept input sequences of up to 8K tokens. Our results reveal that providing long-range context (i.e., beyond the previous 2K tokens) to these models only improves their predictions on a small set of tokens (e.g., those that can be copied from the distant context) and does not help at all for sentence-level prediction tasks. Finally, we discover that PG-19 contains a variety of different document types and domains, and that long-range context helps most for literary novels (as opposed to textbooks or magazines).

1 Introduction

Understanding long documents requires modeling various discourse-level phenomena, including anaphora (Hobbs, 1979; Grosz et al., 1995), argument structure (Grimshaw, 1990), narrative scripts and trajectories (Schank and Abelson, 1977; Labov and Waletzky, 1997), and causal links between concepts (Mooney and DeJong, 1985). Unfortunately, most language models (LMs) are trained to predict the next word given only a small window of local context, which prevents them from using long-range discourse structure to improve their predictions. Many research efforts over the

years have attempted to address this issue: for example, Rosenfeld (1996) incorporated statistics from distant tokens to improve n -gram models, while Ji et al. (2015) added document-level context to neural LMs.

More recently, the Transformer LM (Vaswani et al., 2017), which forms the backbone of state-of-the-art NLP systems (Devlin et al., 2019; Brown et al., 2020), has become the focus of numerous efforts to process longer input sequences. Transformer LMs are constrained by the inefficiency of the self-attention mechanism, whose complexity scales quadratically with the input sequence length. As such, more efficient methods based on sparse attention (Correia et al., 2019) and cached memory (Rae et al., 2020) have been proposed to increase the maximum sequence length, which has progressed from GPT-2’s 1024 tokens (Radford et al., 2019) to 4096 tokens (Zaheer et al., 2020) and finally 8192 tokens (Roy et al., 2021, *Routing Transformer*). When evaluated on the PG-19 benchmark dataset (Rae et al., 2020), which contains long documents in the public domain, these “long-range” Transformer LMs also reach lower perplexities than baseline models on held-out data.

How do “long-range” Transformer LMs make use of the long-range context? Do they actually encode important discourse information to improve their predictions? In this paper, we conduct a series of fine-grained analysis experiments to answer these questions, several of which are inspired by the context analysis of LSTM LMs conducted by Khadelwal et al. (2018). We focus specifically on analyzing the behavior of the state-of-the-art *Routing Transformer* and a simpler baseline model in the presence of various perturbations (e.g., word shuffling, random document replacement), and look closely at how different types of tokens are affected. Our results show that:

- Providing long-range context (i.e., further than 2K tokens away) to these models has *neg-*

ligible impact on the perplexity of tokens near the end of a sequence in aggregate. However, a fine-grained analysis reveals that it does help a small set of tokens (tokens within subword clusters and those that can only be copied from the distant context), as well as particular types of books (fictional and continuous).

- Despite the aforementioned improvements on a fraction of tokens, significantly perturbing the long-term context with word shuffling and random replacement has no notable impact on perplexity overall, suggesting that the evaluated models encode long-range context superficially at best.
- Long-range context is not used for sequence-level prediction tasks that move outside the teacher-forced setting of the previous experiments.

While modern LMs can process much longer input sequences than those of the past, we conclude that they do not exploit the information available in the long-range context. We recommend that future research on long-range LMs includes analysis experiments such as those in our work to shed light on how and when they are using the distant context.

2 Background & Setup

In this section, we first provide an overview of the long-range language models analyzed in this work (Local Transformer, Routing Transformer). Next, we describe the experimental setup used in the remainder of the paper to measure the impact of the long-term context.

2.1 Long-range Language Modeling

Given preceding context tokens $w_{<i}$ (the *prefix*), a language model (LM) computes the probability distribution of the next token $p(w_i | w_{<i})$. LMs are commonly evaluated using perplexity, which is the exponentiated negative log likelihood of a held-out corpus:

$$ppl = \exp \left(-\frac{1}{N} \sum_{i=1}^N \log p(w_i | w_{<i}) \right)$$

Modern LMs are most often implemented with *Transformers* (Vaswani et al., 2017), which compute vector representations for every token in the prefix at multiple layers and combine them together using self-attention. Since self-attention requires

scoring every token in the sequence against every other token, it scales quadratically in both compute time and memory usage, which limits its application to very long sequences.

Local Transformer A simple way to improve the efficiency of self-attention blocks is to constrain the attention at each layer to a local window of the previous k tokens. Such Transformers, which we refer to as *Local Transformers*, can be feasibly scaled up to large input sequence lengths. The receptive field of a Local Transformer scales linearly with the number of layers, as the l^{th} layer of this model can access the previous $k \times l$ tokens (Luong et al., 2015; Child et al., 2019; Sukhbaatar et al., 2019).

Routing Transformer The Routing Transformer (Roy et al., 2021, RT) takes a more intelligent approach to scaling self-attention. Specifically, the RT assigns keys and queries in self-attention to k clusters, the centroids of which are learned during training. A routing attention strategy computes attention A only over the queries Q_i and keys K_j that belong to the same cluster $\mu(Q_i)$ (i.e., those whose centroid is closest to query Q_i).

$$X_i = \sum_{\substack{j: K_j \in \mu(Q_i), \\ j < i}} A_{ij} V_j$$

In contrast to the position-based local attention, this clustering-based attention takes the *content* of the token representations into account. This sparse self-attention mechanism reduces the complexity from $O(N^2)$ to $O(N^{1.5})$ and has led to state-of-the-art results on tasks such as long-form question answering (Krishna et al., 2021).

2.2 Experimental setup

While the previously-described models can be trained with longer inputs than standard Transformers, it remains unclear how they make use of the additional context tokens to improve their predictions of the next word. To shed light on the behavior of long-range Transformer LMs, we perform a series of experiments in which we manipulate the input sequence (both length and content). For token-level experiments (§ 3, § 4), we only evaluate the perplexity of k tokens near the end¹ of an N

¹An artifact exhibited by RT causes tokens at the very end of a sequence to have much higher losses than the others; this

token-long input sequence ($k \ll N$) to focus on the effect of long-range context.²

Dataset: We conduct all of our analyses on the validation set of the PG-19 dataset (Rae et al., 2020). This dataset contains ~ 29 K books from Project Gutenberg repository published before 1919 and was constructed specifically to evaluate long-range LMs (average document length ~ 69 K tokens). The validation set contains 50 books³ and ~ 3 million tokens in total. Evaluating every token in the validation set with large prefix sizes (e.g. 8K tokens) is computationally infeasible.⁴ Thus, we set the number of target tokens per context $k = 10$ and sample a subset of 220K validation tokens for our experiments, which is the same data size used in the LM analysis experiments of Khandelwal et al. (2018). Evaluating RT in this way yields slightly better perplexity on the validation set of PG-19 than using the evaluation setting in the original RT paper (35.2 vs. 36.3). We ensure that the number of tokens sampled from each validation book is proportional to the length of that book.

Models: As training long-range Transformer LMs is also infeasible without immense computational resources, we use publicly-available pre-trained checkpoints for all of our experiments. The Routing Transformer (RT) checkpoint contains 490M parameters and processes sequences up to 8192 tokens long, achieving 33.2 perplexity on the PG-19 test set. The released checkpoint, which was trained on 128 TPUv3 cores for several weeks, has a subword vocabulary of ~ 98 K types,⁵ along with 8 attention heads in each of its 22 layers. The top

phenomena does not exist with LT. After correspondence with the RT authors, we decide to exclude the very last 40 tokens (whose losses are affected) from all of our evaluations. More details about this artifact can be found in the Appendix.

²Language models are normally evaluated on non-overlapping sequences, with the begin and end sequence tokens receiving different amount of context. In our setting, all target tokens have roughly the same amount of context.

³We observe a significant gap of ~ 10 perplexity between the PG-19 test and validation sets and discover that this gap is largely due to the presence of an annotated edition of *The Canterbury Tales and Other Poems*. This book intertwines line-by-line annotations with the main text, which causes the preprocessed version in the dataset to be unreadable. We remove this book in all of our experiments, which decreases the test/validation perplexity gap to ~ 3 .

⁴On one RTX8000 GPU, it takes around 104h to evaluate the entire PG-19 validation set with sequence length 8K and target sequence 10.

⁵We follow the RT paper (Roy et al., 2021) by scaling the loss by 1.248 before computing the perplexity in order to match the word-level perplexity reported by Rae et al. (2020).

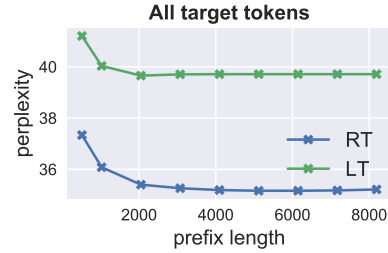


Figure 1: The perplexity of all target tokens plateaus after 2K prefix tokens for both Routing Transformer (RT) and Local Transformer (LT), showing the negligible overall impact of the long-range context.

two RT layers include content-based clustering attention while the remaining are composed entirely of local attention.

Our Local Transformer (LT) is derived from the same checkpoint as RT (and thus has identical model size), except that all clustering heads are replaced with local attention heads. It achieves slightly better perplexity on the PG-19 test set (38.3 vs. 39.3) compared to the LT model trained from scratch by Roy et al. (2021), possibly because the local attention heads learn a better representation of the weight matrices by using the information from the clustering heads.⁶ The attention heads in this model attend to the previous 256 tokens, which results in an effective receptive field of ~ 5 K tokens.⁸ While we would have also liked to analyze other long-range LMs such as the Compressive Transformer (Rae et al., 2020) and Longformer (Beltagy et al., 2020), these models do not have publicly-available PG-19 checkpoints; additionally, they differ from RT and LT in model size, which makes it hard to perform controlled experiments.

3 The effect of longer context

How does the size of the prefix affect the perplexity of long-range Transformer LMs? In this section, we evaluate our RT and LT checkpoints on the

⁶The original fully trained LT checkpoint was not made publicly available before the EMNLP submission deadline.

⁷The RT authors released a new LT checkpoint during the review period of this paper. We also evaluate this newly-released LT checkpoint and include the results in the Appendix F. Both the RT and the LT checkpoints can be found at https://github.com/google-research/google-research/tree/master/routing_transformer

⁸A preliminary experiment verified that the clustering heads in the RT do attend to the long-range context, beyond 5K tokens, demonstrating that it is at least theoretically incorporating more context than the LT.

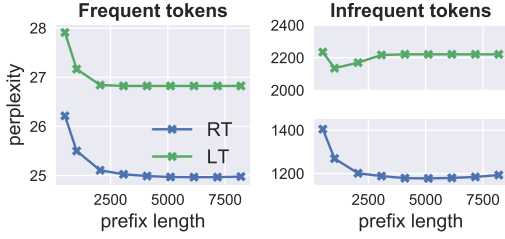


Figure 2: RT perplexity of infrequent tokens continues to decrease until prefix length is 5K.

PG-19 validation set with varied prefix length. We discover that although these models are theoretically able to encode long sequences, increasing the prefix length beyond 2K tokens does not bring discernible improvements in aggregate. However, we do identify small subsets of tokens that benefit from long-range context. Additionally, we find that these models take advantage of long-range context to different degrees on different types of books (e.g., continuous fictional narratives vs. discontinuous magazine articles).

Validation perplexity does not improve when the prefix length grows beyond 2K tokens: As shown in Figure 1, RT perplexity plateaus when evaluated with prefixes longer than 2K.⁹ In contrast, relative to RT, the perplexity curve for the more primitive LT starts flattening earlier at around 1K tokens (note that its effective context size is only 5K). We conclude that while RT’s clustering heads take better advantage of global context than LT, the long-range context beyond 2K tokens is not helping overall. Surprisingly, the perplexity gap between RT and LT is relatively consistent regardless of the prefix length, which indicates that much of RT’s gains do not come from its increased ability to leverage long-range context but rather from better modeling of *local context*.

Infrequent tokens can benefit from increased prefix length: While the overall perplexity does not improve with increasing prefix length, we do observe different behavior when filtering the target tokens by frequency, as shown in Figure 2. We define *frequent* tokens to be the top 10% more frequently-occurring tokens in the subword vocabulary of PG-19 while the rest of tokens are clas-

⁹As a point of comparison, the perplexity of the much smaller LSTM language models evaluated by Khandelwal et al. (2018) plateaus after 200 words. Additionally, Press et al. (2020) discover that the perplexity flattens after 1K for a smaller standard Transformer LM.

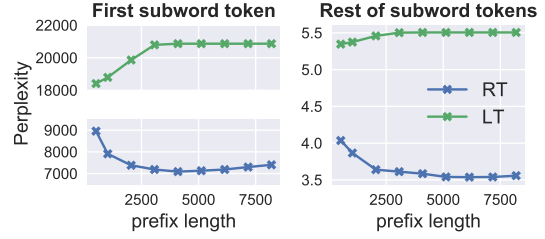


Figure 3: RT perplexity of tokens inside subword clusters continues to decrease until a prefix length of 5K.

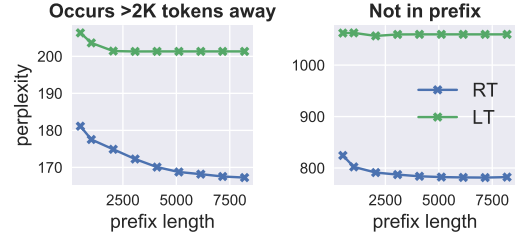


Figure 4: RT perplexity of target tokens whose last appearance is more than 2K tokens away in the prefix keeps decreasing.

sified as *infrequent*.¹⁰ While adding long-range context does not improve either model’s predictions of frequent tokens, the RT’s perplexity of infrequent tokens decreases from ~ 1200 with a prefix length of 2K to 1180 with prefix length of 5K. However, we do observe that infrequent token perplexity increases back to 1200 as the input is further extended, suggesting that the additional context perhaps confounds the model. Meanwhile, the LT is significantly worse than RT on infrequent token prediction, and its perplexity increases as the prefix length is increased.¹¹

Tokens inside a subword cluster benefit from longer contexts: One issue with the previous experiment is that the frequency categorization was computed at a subword level and so may not exactly correspond to word frequency, especially for infrequent words (e.g., entities) that are split into multiple subwords. We therefore do a follow-up experiment by isolating all words that are split into multiple subwords, and then examining perplexity of these tokens as a function of their position in the subword cluster. For example, the word

¹⁰Around 20K tokens in our target token set are classified as infrequent, which amounts to 9% of all target tokens.

¹¹This is likely an artifact due to the elimination of routing attention heads from the RT checkpoint, since the LT trained from scratch does not exhibit such behavior. More details are included in Appendix F.

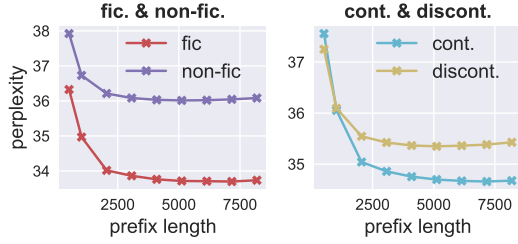


Figure 5: RT takes better advantage of context beyond 2K tokens for fictional and continuous books than non-fictional and discontinuous books, respectively.

“Trocadero” is separated into three subword tokens “Tro”, “cade”, and “ro”. We specifically distinguish between the *first* subword in the cluster (“Tro”) from the *rest* of the subwords (“cade” and “ro”) in the plots shown in Figure 3. The perplexities are computed over 4.1K *first* and 5.1K *rest* subword tokens. The *first* subword category exhibits the same curve shape as those for infrequent tokens for both models, although the magnitude of the perplexities is far higher. The rest of the subwords are far easier for both models to predict, but the RT perplexity curve shows some positive impact from the long-range context until a prefix size of 5K tokens.

Routing Transformers are able to copy tokens that occur in the long-range context: Target subword tokens that can be copied from somewhere in the prefix form another interesting group to analyze.¹² While this is commonplace for frequent words (e.g., determiners, pronouns), it also occurs for entities and rare words (e.g., character names in a novel); sometimes, a word can occur several thousand tokens after its last occurrence. We focus on the latter category of tokens, specifically using a prefix length of 2K tokens as a cutoff to distinguish local and long-range context. Perplexities are computed over 22k tokens which occur last time more than 2K tokens away, and 36K tokens that never appear in the prefix. In particular, the left plot in Figure 4 shows the perplexity of tokens that cannot be found in the previous 2K tokens, but occur somewhere in the long-range context (2K to 8K tokens away). While the LT curve for such tokens plateaus after 2K tokens, indicating that LT cannot take advantage of repeated words in the distant

¹²Note that there is some overlap between the token categories we have analyzed so far. We verify in the Appendix C Table 2 that the overlap between these categories is not significant enough to confound the results.

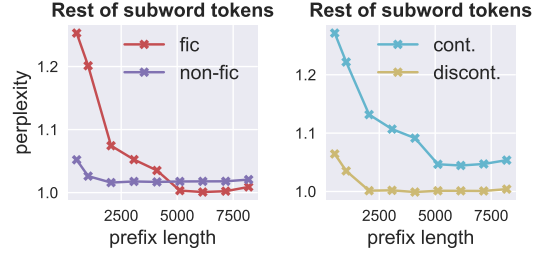


Figure 6: The majority of improvements on tokens inside subword clusters (i.e., excluding the first token in subword clusters) from prefixes longer than 2K tokens comes from fictional and continuous books.

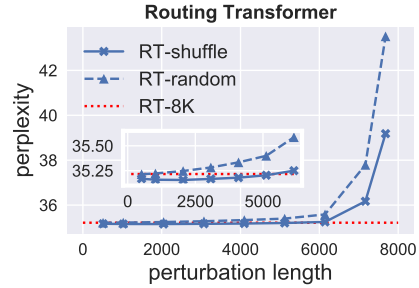


Figure 7: Perturbing up to 6K prefix tokens does not notably affect RT’s overall perplexity. The corresponding plot for LT is included in Appendix D.

context, the RT curve steadily decreases until 8K tokens. The right plot, which shows the subset of target tokens which do not occur anywhere in the short or long-term context, decreases until about 5K tokens and then plateaus. Overall, these results show that long-range context is helpful for tokens that appear even several thousands tokens away.

Following patterns in the long-range context: Besides the token categories examined above, we also qualitatively look at some examples that are too infrequent to analyze at scale. Interestingly, we observe some simple patterns (slightly more complex than copying) that the RT model picks up on. Specifically, it learns to increment chapter numbers even if the previous chapter title appears more than 2K tokens away: for example, when predicting “Chapter V” in the validation book *Keith of the Border*, modifying the previous chapter title “Chapter IV”, which occurs 2300 tokens away, to “Chapter V” causes the loss of the predicted token “V” to increase by over 10.

The impact of book type on the benefits of long-range context: PG-19 contains a diverse array of topics, genres, and formats, not all of which equally

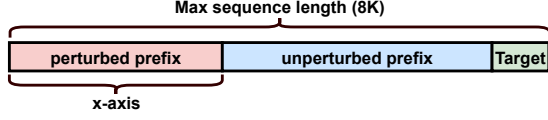


Figure 8: Generic illustration of the x-axis in all the perturbation analysis.

benefit from long-range context modeling. For example, while continuous narratives (e.g., novels) certainly build up many high-level discourse structures over a long sequence of tokens, discontinuous text like magazines, textbooks, or short story collections may require primarily local modeling. To better understand the effect of the type of book on long-range LM perplexity, we annotate every book in PG-19’s validation set as either *fiction* or *non-fiction* and *continuous*¹³ or *discontinuous*.¹⁴ Out of 49 books we annotated, 30 are non-fiction,¹⁵ 31 are discontinuous, and 25 books are both non-fiction and discontinuous.

We observe in Figure 5 that the RT model takes better advantage of long-range context for fictional and continuous books, as the perplexity for these books plateaus at around 5K tokens. Figure 6 shows fictional and continuous books exploit better the long-range context while predicting tokens within subword clusters. Overall, we find the improvement stems largely from continuous and fictional books; more details are included in Appendix B.

4 The perturbation of long-range context

The experiments in the previous section show that incorporating long-range context (further than 2K tokens away from the target) yields only marginal improvements to the overall perplexities of RT and LT. However, the long-range context does have a notable positive impact on a subset of tokens and book types. Do these improvements persist in the presence of severe perturbations to the distant context? If so, this would indicate that they are not encoding any complex discourse structure in the context but rather relying on surface information (e.g., token presence) to make better predictions. In

¹³We consider books with related but distinct sections (such as textbooks) to be discontinuous in our annotation.

¹⁴We also annotate whether the work has been written by the same author or various authors, which is presented in the Appendix.

¹⁵Some magazines contain short stories or poems interspersed with news articles and essays; we count these as non-fiction in our analysis.

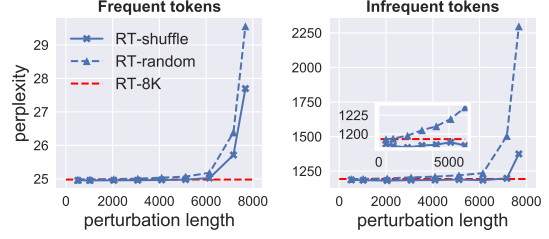


Figure 9: While perturbing up to 6K prefix tokens has little impact on frequent tokens, it increases the perplexity of infrequent tokens.

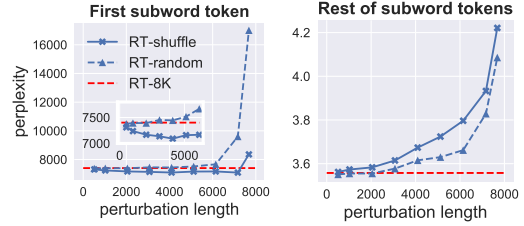


Figure 10: Both shuffling and random replacement increase the perplexity of tokens inside subword clusters, with the former having more negative impact.

this section, we perform a perturbation analysis to quantitatively measure the robustness of the state-of-the-art RT model.¹⁶

Formally, assume we are given a prefix sequence $P = (w_0, w_1, \dots, w_n)$ with which we want to predict target sequence $(w_{n+1}, w_{n+2}, \dots, w_{n+k})$. We apply a perturbation ρ to the first m tokens of the prefix $(w_{0:m})$ to obtain the perturbed prefix

$$\tilde{P} = (\rho(w_0, \dots, w_m), w_{m+1}, \dots, w_n).$$

We define the following three perturbation operations for ρ and report results averaged over five runs for each of them.

- **Sequence shuffling:** Tokens within the perturbed window $w_{0:m}$ are shuffled across the entire window (i.e., sentence boundaries are not respected).
- **Random sequence replacement:** $w_{0:m}$ is replaced with a random sequence from another validation book that is m tokens long.
- **Specific token drop:** Specific tokens within $w_{0:m}$ (e.g., those that occur in the target) are dropped and replaced with the padding token.

¹⁶Figure 20 in the Appendix shows that the Local Transformer never uses context beyond 3K. Due to this limitation, we only present results on RT for in this section.

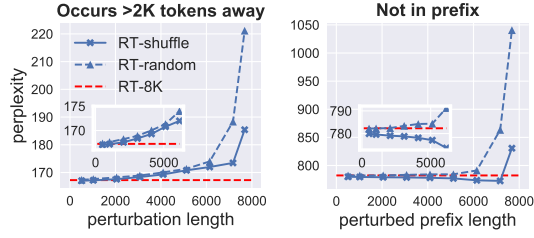


Figure 11: Both perturbation operations increase the perplexity of target tokens whose last appearance in the prefix is more than 2K tokens away.

Sequence-level perturbations further than 2K tokens from the target have minimal impact on perplexity: We first apply sequence-level shuffling and random replacement to the distant context. Both operations have minimal impact on the perplexity of all tokens (Figure 7) as well as frequent/infrequent tokens (Figure 9) provided at least 2K tokens are left unperturbed. However, these perturbations do have increasing impact as the perturbations come closer to the target, especially for infrequent tokens. Zooming in on the long-range context, we find that random replacement consistently results in higher perplexity than shuffling, but also that shuffling distant context actually achieves slightly *lower* perplexity than when the model is given completely unperturbed prefixes. Overall, these results demonstrate that RT is insensitive to the word order of the long-range context.

Tokens inside subword clusters and tokens repeated in the distant context depend on word order: Similar to the analysis in Section 3, the experiments above may hide impacts on small subsets of tokens, which motivates us to do a more fine-grained analysis. We find that tokens inside subword clusters (Figure 10) and those that can only be copied from long-range context (Figure 11) are sensitive to both the order and the content of the remote context. While random replacement is more harmful than shuffling for tokens that can be copied in the distant context (172 shuffled vs 174 random replacement perplexity when perturbing 6K tokens), shuffling is more detrimental for tokens inside subword clusters (3.8 vs 3.7 perplexity when perturbing 6K tokens).

Routing Transformer encodes token identity in the long-range context: While the previous perturbations affected entire contiguous blocks of the prefix, we move now to more targeted perturbations. An interesting question to ask given the observation

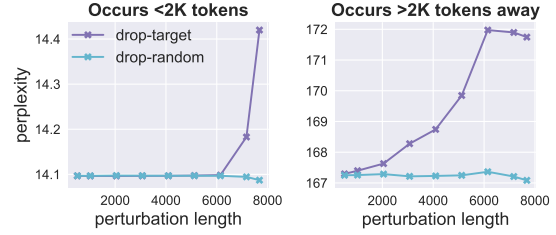


Figure 12: Perplexity of target tokens whose last occurrence in the prefix is within previous 2K tokens (**left**) or more than 2K tokens away (**right**), when dropping either these target tokens or random tokens in the perturbed range. The curve on the right indicates RT memorizes token identity in the distant prefix to some extent.

that RT perplexity decreases on copied tokens as sequence length increases (§ 3) is how much that perplexity decrease depends on word order and surrounding content. In response, we drop tokens in the distant context whose next appearance is in the target sequence. As a control experiment, we drop the same number of random tokens for each perturbation length.

As shown in the right plot of Figure 12, dropping the previous long-range occurrences of target tokens increases the perplexity of those target tokens, which shows that RT indeed memorizes token identity in the long-range context to some extent. The left plot shows that dropping long-range duplicate tokens does not affect tokens that also occur within the local context (i.e., the prior 2K tokens). The flat curve before 6K indicates the model relies only on the most recent occurrences for prediction.

5 Sequence-level analysis

All of the previous experiments have focused on *token-level perplexity*, which is the standard way in which LMs are evaluated. However, the prefixes in these evaluations consist solely of ground-truth text, mirroring the “teacher-forcing” setup that LMs are trained with. When these models are deployed practically to generate text, they have to rely on their previous predictions instead of ground-truth text, and several prior works have noted different behavior in this setting (Wang and Sennrich, 2020; Holtzman et al., 2020; Welleck et al., 2020). In this section, we shift from token-level tasks to analyzing RT and LT performance on *sequence-level* tasks. In particular, we first look at how well the models can memorize an exact sequence in the distant context, as opposed to a single token as we did

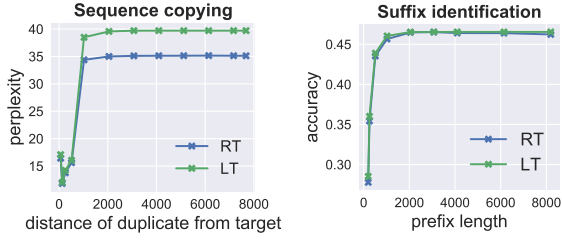


Figure 13: **Left:** Both models assign low perplexity if a duplicate sequence appears within previous 512 tokens. **Right:** Both models have almost identical performance on our suffix identification task. Adding context beyond 2K tokens does not improve performance of either sequence-level prediction task.

Prefix	... If the doctor 's prophecy is correct... (~700 tokens) ... "How far is it to his place?" "Oh, a mile at least. We can have a cab." "A mile?"
Gold suffix:	Then we shall see if there is any truth in what that swab of a doctor said ...
Negative 1:	If I can see Mr.Elberry to-day we may let you have a cheque ...
Negative 2:	It was not until he had signed and sent it off that the full significance of all. . .
Negative 3:	He hurried in, fearing that she might have taken some turn for the worse. . .
Negative 4:	look!!!"Her voice had fallen suddenly to a quivering whisper and she was. . .
Negative 5:	Again the Admiral burst out cheering. "There remains, therefore. . .

Table 1: An example of suffix identification task, the full version of this example is included in Appendix E.

previously. Next, we examine the models' ability to identify which of six 128-token suffixes follows a given prefix, which examines their behavior outside the standard teacher-forced setting.

Sequence-level copying: As a sequence-level analogue to the token-copying analysis in the previous section, we examine both RT and LT's ability to memorize a sequence that occurs in the distant context. To test this ability, we copy the target sequence and paste it into different positions of the prefix. The left plot in Figure 13 shows that both models give a very low perplexity to the target sequence if its duplicate appears within the previous 512 tokens. However, **both models lose their ability to take advantage of the copied sequence if it occurs more than 2K tokens away**. This confirms our previous discovery that sequence order is in general not encoded in the long-range context.

Suffix identification: To move beyond token-level experiments, we adopt a similar setting as the multiple choice task in SWAG (Zellers et al., 2018). Specifically, a prefix is paired with the ground-truth next 128 tokens (or *suffix*) as well as five randomly sampled sequences of length 128 that come from the same book and do not occur in the prefix or gold suffix. We constrain the prefix to end at a full stop and each candidate suffix to start from a new sentence so that the difference in perplexity is not due to ungrammaticality. An example is shown in Table 1. We construct 7K examples and compute the accuracy of both models at correctly choosing the correct suffix. The model makes a correct prediction when the gold suffix has lower perplexity than all other suffixes. As shown in the right plot of Figure 13, increasing prefix length beyond 2K does not improve suffix identification accuracy. Surprisingly, the LT and RT model have almost identical (and poor) performance on this task.¹⁷ While RT is a significantly better LM in terms of token-level perplexity, it does not appear to be superior in terms of using long-range context to improve sequence prediction. Overall, both models often predict obviously wrong negative suffixes: the full version of Table 1 together with RT's perplexity score for each suffix is included in Appendix E.

Combined with our previous token-level analysis, we conclude that the distant context helps a subset of tokens in superficial ways; however, distant context is currently not helpful for sequence-level prediction tasks.

6 Related work

Our work examines recent advances in efficient Transformer variants (Sukhbaatar et al., 2019; Kitaev et al., 2020; Choromanski et al., 2021; Tay et al., 2021; Katharopoulos et al., 2020; Wang et al., 2020; Wu et al., 2020) that accept longer sequences than prior approaches. Longer effective context size is often achieved by sparse attention (Child et al., 2019), recurrence (Dai et al., 2019), and cached memory (Weston et al., 2015; Rae et al., 2020). Our work is also related to methods that incorporate long context (Wang and Cho, 2016) as well as document-level tasks that inherently require modeling long-range context (Zhang et al., 2018;

¹⁷While evaluating on the newly released LT checkpoint, the performance of LT is slightly worse, but the trend is similar. Adding context beyond 2K tokens does not keep improving the suffix identification accuracy. We direct reader to Appendix F for more details.

Hofstätter et al., 2020; Zhang et al., 2020).

This work is also similar to other analysis of language models, especially for long-range context. Khandelwal et al. (2018) analyze the usage of long-term context of smaller LSTM LMs. Sharan et al. (2018) prove long-term context is not needed for HMM LM due to teacher forcing. Rae and Razavi (2020) conduct an analysis exclusively for the Transformer-XL (Dai et al., 2019) model. Rae et al. (2020) show that Compressive Transformer improves the performance of infrequent tokens. Our work also relates to that of Lai et al. (2020), who investigate the impact of context for pretrained masked LMs. More recently, Press et al. (2020) also observe negligible benefits of long-term context; we step further in this direction by exploring larger models with more fine-grained analysis.

7 Conclusion

We perform a fine-grained analysis of the impact of long-range context to both token- and sequence-level improvements on two long-range Transformer language models, using the PG-19 dataset as a testbed. Our results suggest these models rarely take advantage of the long-term context, and when they do it is mostly in superficial ways (e.g. by copying rare tokens from far away). With the proliferation of research in increasing the input size of Transformer LMs, we hope that our research will lead to more meaningful progress on integrating discourse information into these models.

Ethical concerns

Misuse of language models The two large language models we evaluated in this work share common ethical concerns with works on language models and language generation. These pre-trained LMs can be used maliciously to generate unfaithful, hallucinated, and biased output. Our reported results do not include any kind of generation.

Energy costs We conduct all our analysis experiments on RTX8000 GPUs. Although our work does not include training large language models, the energy costs of evaluating large pre-trained LMs, such as the Routing Transformer, should not be ignored. Each example of 8K tokens long takes around 1.3s \sim 1.4s to run one forward pass with the RT model. We hope our analysis can shed light on more efficient and effective method to encode long-term context.

Acknowledgements

We are grateful to Aurko Roy for releasing the code and checkpoints and for discussing the Routing Transformer results with us. We thank Nader Akoury, Brendan O’Connor and the rest of UMass NLP group for the great advice on the draft of this paper. We thank the anonymous reviewers for their thoughtful comments on the paper. This project was partially funded by a grant from Intuit AI and also by award IIS-1955567 from the National Science Foundation (NSF).

References

- Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. [Longformer: The long-document transformer](#).
- Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
- Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. [Generating long sequences with sparse transformers](#).
- Krzysztof Marcin Choromanski, Valerii Likhoshesterov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Beller, Lucy J Colwell, and Adrian Weller. 2021. [Rethinking attention with performers](#). In *International Conference on Learning Representations*.
- Gonalo M. Correia, Vlad Niculae, and Andr  F. T. Martins. 2019. [Adaptively sparse transformers](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2174–2184, Hong Kong, China. Association for Computational Linguistics.
- 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.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [Bert: Pre-training of deep bidirectional transformers for language understanding](#).
- Jane Grimshaw. 1990. *Argument structure*. the MIT Press.

- Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. [Centering: A framework for modeling the local coherence of discourse](#). *Computational Linguistics*, 21(2):203–225.
- Jerry R Hobbs. 1979. Coherence and coreference. *Cognitive science*, 3(1):67–90.
- Sebastian Hofstätter, Hamed Zamani, Bhaskar Mitra, Nick Craswell, and Allan Hanbury. 2020. [Local self-attention over long text for efficient document retrieval](#).
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. [The curious case of neural text de-generation](#). In *International Conference on Learning Representations*.
- Yangfeng Ji, Trevor Cohn, Lingpeng Kong, Chris Dyer, and Jacob Eisenstein. 2015. Document context language models. *arXiv preprint arXiv:1511.03962*.
- 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*.
- Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. 2018. [Sharp nearby, fuzzy far away: How neural language models use context](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 284–294, Melbourne, Australia. Association for Computational Linguistics.
- Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2020. [Reformer: The efficient transformer](#). In *International Conference on Learning Representations*.
- Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *North American Association for Computational Linguistics*.
- William Labov and Joshua Waletzky. 1997. Narrative analysis: Oral versions of personal experience.
- Yi-An Lai, Garima Lalwani, and Yi Zhang. 2020. [Context analysis for pre-trained masked language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3789–3804, Online. Association for Computational Linguistics.
- Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. [Effective approaches to attention-based neural machine translation](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.
- Raymond J Mooney and Gerald DeJong. 1985. Learning schemata for natural language processing. In *IJCAI*, pages 681–687.
- Ofir Press, Noah A. Smith, and Mike Lewis. 2020. [Shortformer: Better language modeling using shorter inputs](#).
- 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 Rae and Ali Razavi. 2020. [Do transformers need deep long-range memory?](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7524–7529, Online. Association for Computational Linguistics.
- 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*.
- Roni Rosenfeld. 1996. A maximum entropy approach to adaptive statistical language modeling.
- 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.
- Roger C Schank and Robert P Abelson. 1977. Scripts, plans, goals, and understanding: an inquiry into human knowledge structures.
- Vatsal Sharan, Sham Kakade, Percy Liang, and Gregory Valiant. 2018. [Prediction with a short memory](#). In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018*, page 1074–1087, New York, NY, USA. Association for Computing Machinery.
- Sainbayar Sukhbaatar, Edouard Grave, Piotr Bojanowski, and Armand Joulin. 2019. [Adaptive attention span in transformers](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 331–335, Florence, Italy. Association for Computational Linguistics.
- Yi Tay, Dara Bahri, Donald Metzler, Da-Cheng Juan, Zhe Zhao, and Che Zheng. 2021. [Synthesizer: Re-thinking self-attention for transformer models](#).
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Chaojun Wang and Rico Sennrich. 2020. [On exposure bias, hallucination and domain shift in neural machine translation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552, Online. Association for Computational Linguistics.

- Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. [Linformer: Self-attention with linear complexity](#).
- Tian Wang and Kyunghyun Cho. 2016. [Larger-context language modelling with recurrent neural network](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1319–1329, Berlin, Germany. Association for Computational Linguistics.
- Sean Welleck, Ilya Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. [Neural text generation with unlikelihood training](#). In *International Conference on Learning Representations*.
- Jason Weston, Sumit Chopra, and Antoine Bordes. 2015. [Memory networks](#).
- Zhanghao Wu, Zhijian Liu, Ji Lin, Yujun Lin, and Song Han. 2020. [Lite transformer with long-short range attention](#).
- 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.
- Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. [Improving the transformer translation model with document-level context](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542, Brussels, Belgium. Association for Computational Linguistics.
- Pei Zhang, Boxing Chen, Niyu Ge, and Kai Fan. 2020. [Long-short term masking transformer: A simple but effective baseline for document-level neural machine translation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1081–1087, Online. Association for Computational Linguistics.

A Routing Transformer and end-sequence degradation

In our analysis, instead of picking the last 10 tokens in a sequence, we chose the last 50 to last 40 tokens due to an artifact introduced by the clustering heads in the RT model. We find that in general the last 20 tokens in a sequence tend to have increasing perplexity as we evaluate on longer and longer sequence lengths. As shown in Figure 14, this phenomenon is only native to the RT model and disappears when the clustering heads are replaced with local attentions. Therefore, to make the RT and the LT comparable, we select the tokens from the range that is not affected by the end-sequence issue. Although it’s not the last 10 tokens, this short target chunk is still located near the end of a sequence, preceded by enough long context.

B Effect of longer context

In section 3 we discussed that books that are fictional and continuous benefit more from the long-range context. We also annotated the validation set by the authorship (i.e., whether a book is written by single author or various authors). Out of 49 books, 11 are written by various authors, 10 of which are non-fictions. Due to this overlap, we only show results of fic/non-fic in the main text.

In this section, we also further break down all targets to the three types of tokens we examined in section 3, and display the results by book types. Perplexity of infrequent tokens (Figure 15), tokens inside subword clusters (Figure 16), and tokens whose last occurrence is more than 2K tokens away (Figure 17). In general, for the small set of tokens whose perplexity keep decreasing as adding in more context, the major source of improvements are from the continuous and fictional books.

	infreq	in-subword	distant
infreq	1.	0.09	0.08
in-subword	0.36	1.	0.1
distant	0.07	0.02	1.

Table 2: Ratio of overlapped target tokens among different categorizations. **infreq** are infrequent tokens, **in-subword** are tokens within a subword cluster (i.e., excluding the first word in a subword cluster), **distant** are tokens only appear in the distant context (more than 2K away). Row 1 column 3 the number 0.08 indicates around 8% of the infrequent tokens are those that can only be found in the distant context.

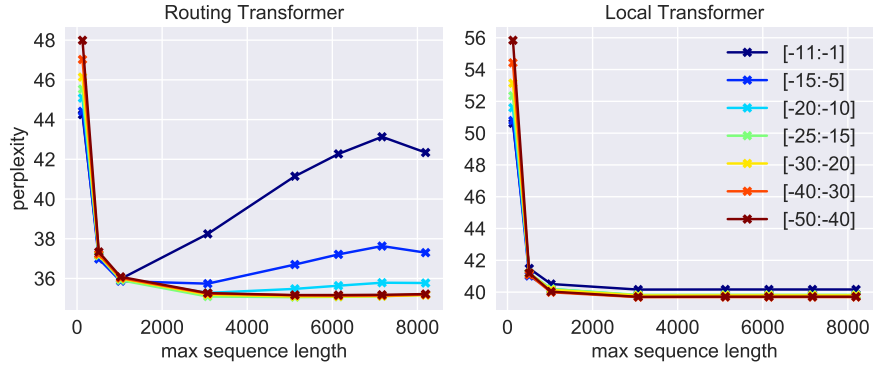


Figure 14: Perplexity of target chunk of 10 tokens long near the end of sequence. The legend indicates how far away the evaluated chunk is from the end of sequence (e.g., [-15:-5] means evaluating the last 15 tokens to the last 5 tokens). The left plot shows the clustering mechanism in the Routing Transformer assigns higher perplexity to around the last 15 tokens in a sequence. Due to this artifacts, in our analysis, we avoid tokens within this range to make comparable comparison with the Local Transformer. In the main text, our analysis are conducted over the last 50 to last 40 tokens, which are not affected by this artifact.

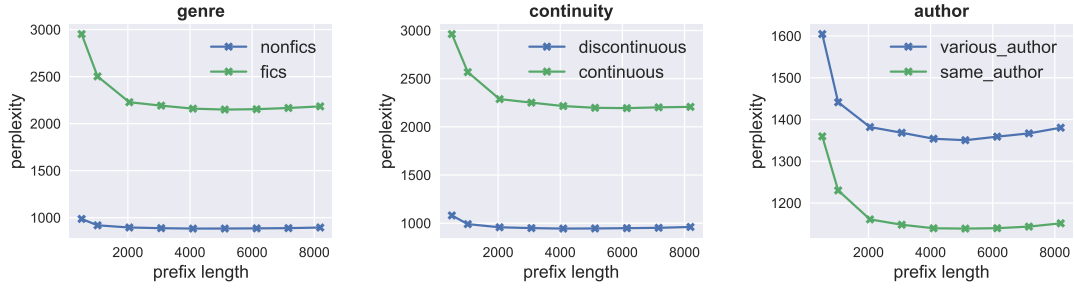


Figure 15: Perplexity of infrequent target tokens, broken down by genre (**left**), continuity (**middle**), and authorship (**right**). Perplexity of infrequent tokens in fictional, continuous and single-authored books decreases as the context length increases to around 5K. On the other hand, the rest types of books rely on more local context while predicting the infrequent tokens.

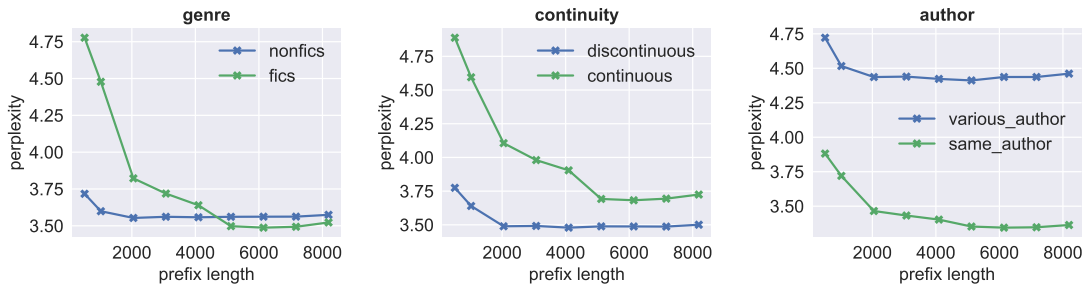


Figure 16: Perplexity of target tokens inside subword clusters (i.e., excluding the first subword in each cluster), broken down by genre (**left**), continuity (**middle**), and authorship (**right**). Perplexity of these tokens in fictional and continuous books improves as the length increases up to around 4K, whereas nonfictional and discontinuous books are not taking any advantage of long-range context at all.

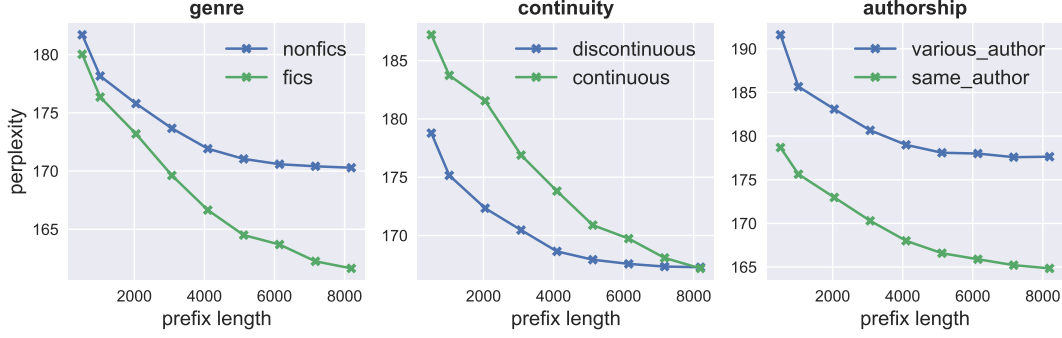


Figure 17: Perplexity of target tokens that can only be found in the distant context, when evaluated with the Routing Transformer on subset of PG-19 validation set, broken down by genre (**left**), continuity (**middle**), and authorship (**right**). Fictional, continuous, and single-authored books continue to have improved perplexity for this type of tokens as the prefix length increases up to 8K. The single-authored books also contain discontinuous books which require less modeling of long-range context. The decreasing curve indicates the model might have acquired author specific token statistics from incorporating longer context.

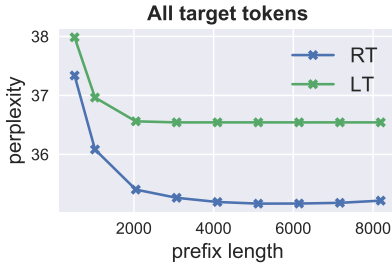


Figure 18: The perplexity of all target tokens plateaus after 2K prefix tokens for both Routing Transformer and Local Transformer. The LT model is the one released in 2021 summer.

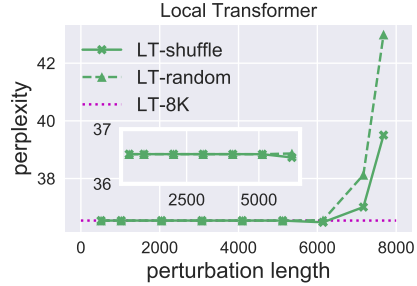


Figure 19: Perplexity of all target tokens when evaluated with Local Transformer released in 2021 summer.

C Token overlaps

We have shown in section 3 that infrequent tokens, tokens inside subword clusters, and tokens that can only be copied from distant context benefit from context longer than 2K tokens. It is possible that these improvements come from the same set of tokens shared across these three types of tokens. To verify if there are significant overlaps among those three types of tokens, we compute the overlapped ratio in Table 2. Except for in-subword and infrequent tokens, the overlapped ratios are all below 0.1.

D Perturbation

Perplexity with perturbed distant prefix when evaluated with Local Transformer is shown in Figure 20. Perplexity hardly changes when perturbing up to around 6K tokens. Because LT doesn't properly use long-range context, we only present the results

of Routing Transformer in the main text.

E Suffix Identification

In the main text, we present the suffix identification results with 128-token long suffix. Here, we provide results when evaluate the accuracy with suffix of different length. Interestingly, the accuracy of distinguishing a gold suffix first increases with the suffix length, reaching the best when the suffix length is around 10 to 20, then decreases as the suffix becomes longer. This is likely because the sequence becomes more probable as incorporating more local context (part of the suffix).

In Table 3 and Table 4, we present a complete example of the suffix identification task. The prefix contains 1024 tokens, and each of the suffixes contains 128 tokens. Lower perplexity of obviously wrong suffix (e.g. negative 1,3,4) indicates current RT model is not properly taking advantage of distant context to make sequence-level predictions.

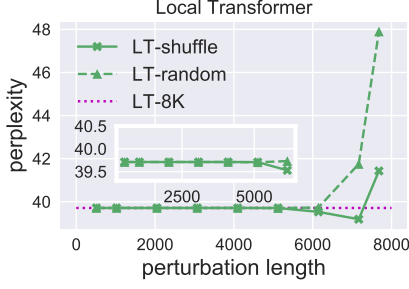


Figure 20: Perplexity of all target tokens when evaluated with Local Transformer derived from the RT checkpoint.

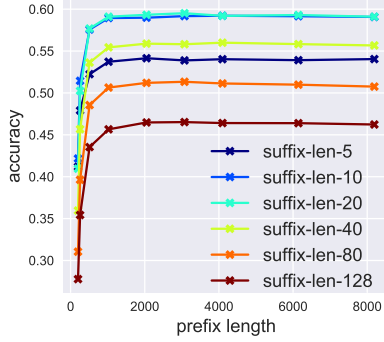


Figure 21: Suffix identification accuracy when evaluated with different prefix length.

F Local Transformer checkpoint results

In the main text, we analyzed both the RT checkpoint and an LT model derived from the same RT checkpoint by replacing the clustering heads with local attention heads. After the submission deadline and before the camera ready deadline, the author of the Routing Transformer released a new LT checkpoint, which has 24 layers in total¹⁸. To make sure the behavior of our former LT is the same with an LT trained from scratch, we also conducted all our analysis again on this new checkpoint. Overall, we find the new LT checkpoint performs slightly better on token-level tasks but is inferior to the previous LT checkpoint on suffix identification, a sequence-level task. Since the trends are the same, no conclusion needs to be changed.

The Effect of Longer Context In Figure 18 are the perplexities in aggregate over all target tokens of both the RT and the new LT checkpoints. The target tokens are the same ones we used in the main text. The new LT has better perplexity than the one

¹⁸Both the RT model and the *former* LT model have 22 layers.

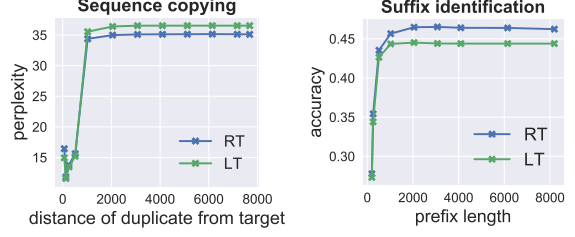


Figure 22: **Left:** Both models assign low perplexity if a duplicate sequence appears within previous 512 tokens. **Right:** Adding context beyond 2K tokens does not improve performance of suffix identification. Moreover, the recently released LT performs worse than the one derived from the RT checkpoint.

presented in the main text (36.5 vs. 40 as the prefix length extends to 8K).

Perturbation of long-range context Similar to the former LT (Figure 20), the newly released LT checkpoint (Figure 19) is not sensitive at all to the perturbation further than local 2K tokens. Both models are impacted by local random replacement more than shuffling, however, the new LT checkpoint has overall better perplexity than the RT-derived LT.

Sequence-level tasks Figure 22 shows the performance of both RT and the released LT checkpoint on sequence-level tasks. Compared to the results in the main text, the new LT checkpoint performs better at sequence-copying task, however, the trend remains the same – the order of tokens beyond 2K tokens is not properly encoded. On the other hand, the new LT checkpoint is slightly worse in suffix identification while the former RT-derived LT has almost identical performance as the RT. This implies even though the clustering heads are removed from the previous LT, useful information is preserved by the local attention heads.

Overall, we find the released LT checkpoint has better token-level performance but performs worse on suffix identification. Each plot shares the same trend with its corresponding one in the main text, thus no conclusion needs to be modified.

Prefix	<p>Not a bit, sir.Out with it!I have faced death too often to flinch from it now, though I saw it as near me as you are.""Well, well, we must go by averages of course.Shall we say two years?I should think that you have a full two years before you.""In two years your pension would bring you in L1,600.Now I will do my very best for you, Admiral!I will advance you L2,000, and you can make over to me your pension for your life.It is pure speculation on my part.If you die to-morrow I lose my money.If the doctor's prophecy is correct I shall still be out of pocket.If you live a little longer, then I may see my money again.It is the very best I can do for you.""Then you wish to buy my pension?""Yes, for two thousand down.""And if I live for twenty years?""Oh, in that case of course my speculation would be more successful.But you have heard the doctor's opinion.""Would you advance the money instantly?""You should have a thousand at once.The other thousand I should expect you to take in furniture.""In furniture?""Yes, Admiral.We shall do you a beautiful houseful at that sum.It is the custom of my clients to take half in furniture."The Admiral sat in dire perplexity.He had come out to get money, and to go back without any, to be powerless to help when his boy needed every shilling to save him from disaster, that would be very bitter to him.On the other hand, it was so much that he surrendered, and so little that he received.Little, and yet something.Would it not be better than going back empty-handed?He saw the yellow backed chequebook upon the table.The moneylender opened it and dipped his pen into the ink."Shall I fill it up?"said he."I think, Admiral," remarked Westmacott, "that we had better have a little walk and some luncheon before we settle this matter.""Oh, we may as well do it at once.It would be absurd to postpone it now," Metaxa spoke with some heat, and his eyes glinted angrily from between his narrow lids at the imperturbable Charles.The Admiral was simple in money matters, but he had seen much of men and had learned to read them.He saw that venomous glance, and saw too that intense eagerness was peeping out from beneath the careless air which the agent had assumed."You're quite right, Westmacott," said he."We'll have a little walk before we settle it.""But I may not be here this afternoon.""Then we must choose another day.""But why not settle it now?""Because I prefer not," said the Admiral shortly."Very well.But remember that my offer is only for to-day.It is off unless you take it at once.""Let it be off, then.""There's my fee," cried the doctor."How much?""A guinea."The Admiral threw a pound and a shilling upon the table."Come, Westmacott," said he, and they walked together from the room."I don't like it," said Charles, when they found themselves in the street once more;"I don't profess to be a very sharp chap, but this is a trifle too thin.What did he want to go out and speak to the doctor for?And how very convenient this tale of a weak heart was!I believe they are a couple of rogues, and in league with each other.""A shark and a pilot fish," said the Admiral."I'll tell you what I propose, sir.There's a lawyer named McAdam who does my aunt's business.He is a very honest fellow, and lives at the other side of Poultry.We'll go over to him together and have his opinion about the whole matter.""How far is it to his place?""Oh, a mile at least.We can have a cab.""A mile?</p>
Gold suffix (ppl=74.4)	<p>Then we shall see if there is any truth in what that swab of a doctor said.Come, my boy, and clap on all sail, and see who can stay the longest."Then the sober denizens of the heart of business London saw a singular sight as they returned from their luncheons.Down the roadway, dodging among cabs and carts, ran a weather-stained elderly man, with wide flapping black hat, and homely suit of tweeds.With elbows braced back, hands clenched near his armpits, and chest protruded, he scudded along, while close at his heels lumbered a large-limbed, heavy, yellow mustached young man, who seemed to feel the exercise a good deal more than his senior.On they dashed, helter-skelter, until they pulled up panting at the office where the lawyer of the</p>

Table 3: Suffix identification example, extracted from the book *Beyond the City* by Arthur Conan Doyle. For each suffix, we show the perplexity of the suffix evaluated with the Routing Transformer. RT assigns a lot lower perplexity to examples (e.g. negative 1,3,4) that are obviously wrong given the prefix.

Negative 1 (ppl=34.6):	If I can see Mr.Elberry to-day we may let you have a cheque to-morrow.Try another pinch.No?Well, good-bye.I am very happy to have been of service."Mr.McAdam bowed them out, for he was a very busy man, and they found themselves in the street once more with lighter hearts than when they had left it."Well, Westmacott, I am sure I am very much obliged to you," said the Admiral."You have stood by me when I was the better for a little help, for I'm clean out of my soundings among these city sharks.But I've something to do now which is more in my own line, and I need not trouble you any more.""Oh, it is no trouble.I have nothing
Negative 2 (ppl=71.69):	It was not until he had signed and sent it off that the full significance of all that he had done broke upon him.He had sacrificed everything.His pension was gone.He had nothing save only what he could earn.But the stout old heart never quailed.He waited eagerly for a letter from the Saint Lawrence Shipping Company, and in the meanwhile he gave his landlord a quarter's notice.Hundred pound a year houses would in future be a luxury which he could not aspire to.A small lodging in some inexpensive part of London must be the substitute for his breezy Norwood villa.So be it, then!Better that a thousand fold than that his name should be associated with failure and disgrace.On that morning Harold Denver was to meet the creditors of the firm, and to explain the situation to them.It was a hateful task
Negative 3 (ppl=33.54):	He hurried in, fearing that she might have taken some turn for the worse, but he was reassured to find her sitting up in her bed, with Clara and Ida Walker in attendance upon her.She had removed the handkerchief, and had put on a little cap with pink ribbons, and a maroon dressing-jacket, daintily fulled at the neck and sleeves."My dear friend," said she as he entered, "I wish to make a last few remarks to you.No, no," she continued, laughing, as she saw a look of dismay upon his face."I shall not dream of dying for at least another thirty years.A woman should be ashamed to die before she is seventy.I wish, Clara, that you would ask your father to step up.And you, Ida, just pass me my cigarettes, and
Negative 4 (ppl=34.92):	look!!!"Her voice had fallen suddenly to a quivering whisper and she was pointing to the Westmacotts' house.Her sister gave a gasp of horror, and stood with a clutch at Monica's arm, staring in the same direction.There was a light in the front room, a slight, wavering light such as would be given by a small candle or taper.The blind was down, but the light shone dimly through.Outside in the garden, with his figure outlined against the luminous square, there stood a man, his back to the road, his two hands upon the window ledge, and his body rather bent as though he were trying to peep in past the blind.So absolutely still and motionless was he that in spite of the moon they might well have overlooked him were it not for that tell-tale light behind.
Negative 5 (ppl=47.66):	Again the Admiral burst out cheering."There remains, therefore, about L3,200 which has to be found within ten days.No man shall lose by me.I gave them my word in the room that if I worked my soul out of my body every one of them should be paid.I shall not spend a penny upon myself until it is done.But some of them can't wait.They are poor men themselves, and must have their money.They have issued a warrant for Pearson's arrest.But they think that he has got away to the States.""These men shall have their money," said the Admiral."Dad!""Yes, my boy, you don't know the resources of the family.One never does know until one tries.What have you yourself now?" "I have about a

Table 4: Suffix identification example, extracted from the book *Beyond the City* by Arthur Conan Doyle. For each suffix, we show the perplexity of the suffix evaluated with the Routing Transformer. RT assigns a lot lower perplexity to examples (e.g. negative 1,3,4) that are obviously wrong given the prefix.