Hyperparameter	PTB	Wiki
Word Emb. Size	400	400
Hidden State Dim	1150	1150
Layers	3	3
Optimizer	ASGD	ASGD
Learning Rate	30	30
Gradient clip	0.25	0.25
Epochs (train)	500	750
Epochs (finetune)	500 (max)	750 (max)
Batch Size	20	80
Sequence Length	70	70
LSTM Layer Dropout	0.25	0.2
Recurrent Dropout	0.5	0.5
Word Emb. Dropout	0.4	0.65
Word Dropout	0.1	0.1
FF Layers Dropout	0.4	0.4
Weight Decay	1.2×10^{-6}	1.2×10^{-6}

Table 2: Hyperparameter Settings.

A Hyperparameter settings

We train a vanilla LSTM language model, augmented with dropout on recurrent connections, embedding weights, and all input and output connections (Wan et al., 2013; Gal and Ghahramani, 2016), weight tying between the word embedding and softmax layers (Inan et al., 2017; Press and Wolf, 2017), variable length backpropagation sequences and the averaging SGD optimizer (Merity et al., 2018). We provide the key hyperparameter settings for the model in Table 2. These are the default settings suggested by (Merity et al., 2018).

B Additional Figures

This section contains all figures complementary to those presented in the main text. Some figures, such as Figures 1b, 1d etc. present results for only one of the two datasets, and we present the results for the other dataset here. It is important to note that the analysis and conclusions remain unchanged. Just as before, all results are averaged from three models trained with different random seeds. Error bars on curves represent the standard deviation and those on bar charts represent 95% confidence intervals.

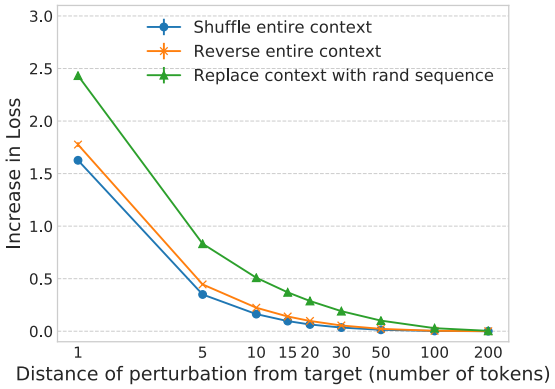


Figure 8: Complementary to Figure 2b. Perturb global order, i.e. all tokens in the context before a given point, in PTB. Effects of shuffling and reversing the order of words in 300 tokens of context, relative to an unperturbed baseline. Changing the global order of words within the context does not affect loss beyond 50 tokens.

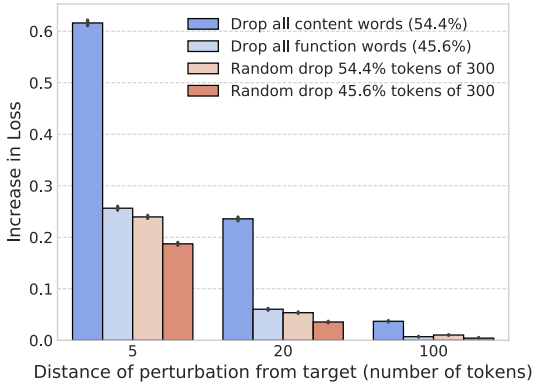
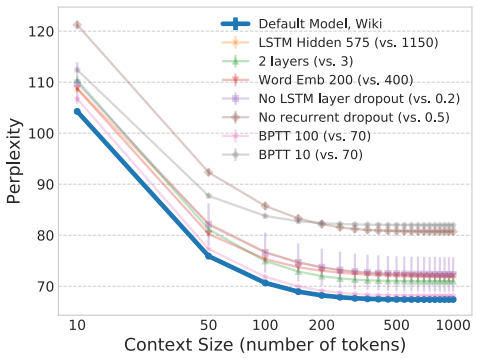
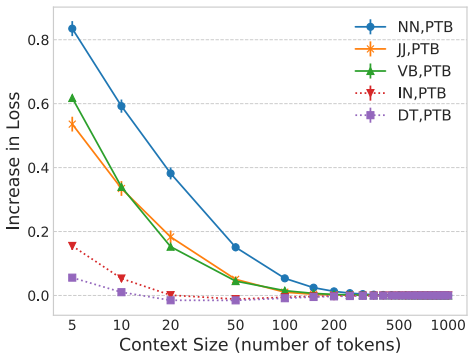


Figure 9: Complementary to Figure 3. Effect of dropping content and function words from 300 tokens of context relative to an unperturbed baseline, on Wiki. Dropping both content and function words 5 tokens away from the target results in a nontrivial increase in loss, whereas beyond 20 tokens, content words are far more relevant.

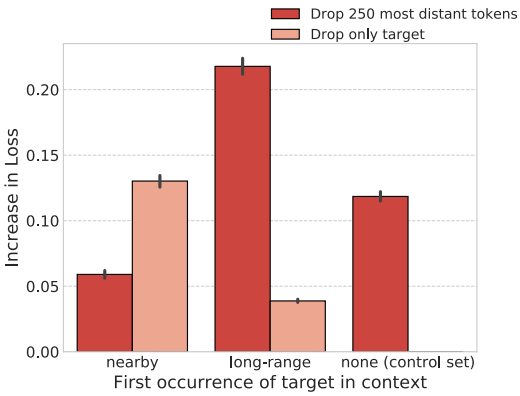


(a) Changing model hyperparameters for Wiki.

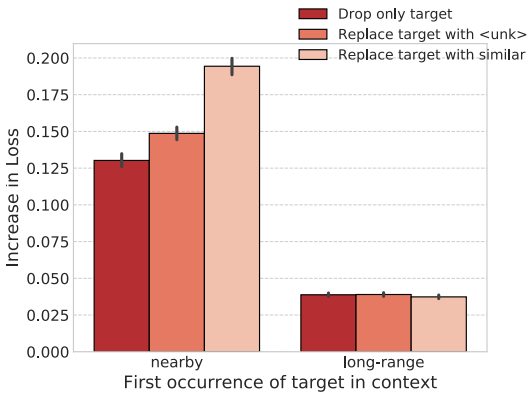


(b) Different parts-of-speech for PTB.

Figure 10: Complementary to Figures 1b and 1d, respectively. Effects of varying the number of tokens provided in the context, as compared to the same model provided with infinite context. Increase in loss represents an absolute increase in NLL over the entire corpus, due to restricted context. **(a)** Changing model hyperparameters does not change the context usage trend, but does change model performance. We report perplexities to highlight the consistent trend. **(b)** Content words need more context than function words.



(a) Dropping tokens



(b) Perturbing occurrences of target word in context.

Figure 11: Complementary to Figure 4. Effects of perturbing the target word in the context compared to dropping long-range context altogether, on Wiki. **(a)** Words that can only be copied from long-range context are more sensitive to dropping all the distant words than to dropping the target. For words that can be copied from nearby context, dropping only the target has a much larger effect on loss compared to dropping the long-range context. **(b)** Replacing the target word with other tokens from vocabulary hurts more than dropping it from the context, for words that can be copied from nearby context, but has no effect on words that can only be copied from far away.

) . Standing roughly 15 metres (49 ft) away , the cadres now raised their weapons . " You have taken our land , " one of them said . " Please don 't shoot us ! " one of the passengers cried , just before they were killed by a sustained burst of automatic gunfire . </s> Having collected water from the nearby village , UNK and his companions were almost back at the crash site when they heard the shots . UNK it was personal ammunition in the luggage exploding in the heat , they continued on their way , and called out to the other passengers , who they thought were still alive . This alerted the insurgents to the presence of more survivors ; one of the guerrillas told UNK 's group to " come here " . The insurgents then opened fire on their general location , prompting UNK and the others to flee . Hill and the UNK also ran ; they revealed their positions to the fighters in their UNK , but successfully hid themselves behind a ridge . After Hill and the others had hidden there for about two **hours**

Figure 12: Failure of neural cache on Wiki. Lightly shaded regions show flat distribution.

La Fortuna , Mexico . UNK just off the coast of Mexico , the system interacted with land and began weakening . UNK later , convection rapidly diminished as dry air became entrained in the circulation . In response to quick degradation of the system 's structure , the NHC downgraded UNK to a tropical storm . Rapid weakening continued throughout the day and by the evening hours , the storm no longer had a defined circulation . Lacking an organized center and deep convection , the final advisory was issued on UNK . The storm 's remnants persisted for several more hours before dissipating roughly 175 mi (280 km) southwest of Cabo Corrientes , Mexico . </s> </s> = = Preparations and impact = = </s> </s> Following the classification of Tropical Depression Two @-@ E on June 19 , the Government of Mexico issued a tropical storm warning for coastal areas between UNK and Manzanillo . A hurricane watch was also put in place from UNK de UNK to Punta San UNK . Later that day , the tropical storm warning was upgraded to a hurricane warning and the watch was extended westward to La **Fortuna**

Figure 13: Success of neural cache on Wiki. Brightly shaded region shows peaky distribution.