Appendix A
Language Modeling Head component activations

Due to space constraints, we will describe some of the experiments in this section. Figure 4 is similar to Figure 2, but lists all of the LM head components. Reiterating, we consider a single input token sequence, where one of the tokens is '[MASK]'; to find the contribution of individual LM head components to the token reconstruction ability, we consequently truncate the LM head *encoder* at each component, computing the inner products between (1) output of the truncated models and (2) word embeddings; thus getting the vocabulary token scores. Then we compare these scores to the scores produced by the entire LM head, sorting both score vectors by the elements of the score vector produced by the truncated model. Since we can truncate the model (1) at the transformer stack, (2) at the dense weights in LM head, (3) at the dense bias in LM head, (4) at GELU in LM head, (5) at LayerNorm normalization in LM head, (6) at LayerNorm elementwise weight multiplication and (7) at LayerNorm elementwise bias addition, we have a total of 7 plots here. Since we sort one sequence by elements of the other one, there are 2 lines in each plot, the blue monotonic line corresponding to the sequence used for sorting, and orange one corresponding to the sorted sequence. We also vary this experiment, swapping the sorted sequences (Figure 6 contains a column of LM head outputs sorted by component-produced outputs on the left, and a column of component-produced outputs sorted by LM head outputs on the right), and using transformer stack output instead of LM head output to produce a pair of columns in Figure 5.

Figure 7 decomposes Figure 3 into distinct plots for each sample perturbation strategy (final '.' removal, '[SEP]' removal, replacing of the '[MASK]' token), increasing the visibility of differences between the plots but also taking much more space.

Table V contains the pairwise Spearman's rank-order correlations for the sequences plotted in Figure 4. The last line of the table contains pairwise correlations between LM head output and sequences produced by each of the truncated models; we can see a significant correlation increase after the LayerNorm normalization layer, becoming even more apparent after the elementwise multiplication. Figure 8 and Figure 9 contain LayerNorm weights and biases accordingly.

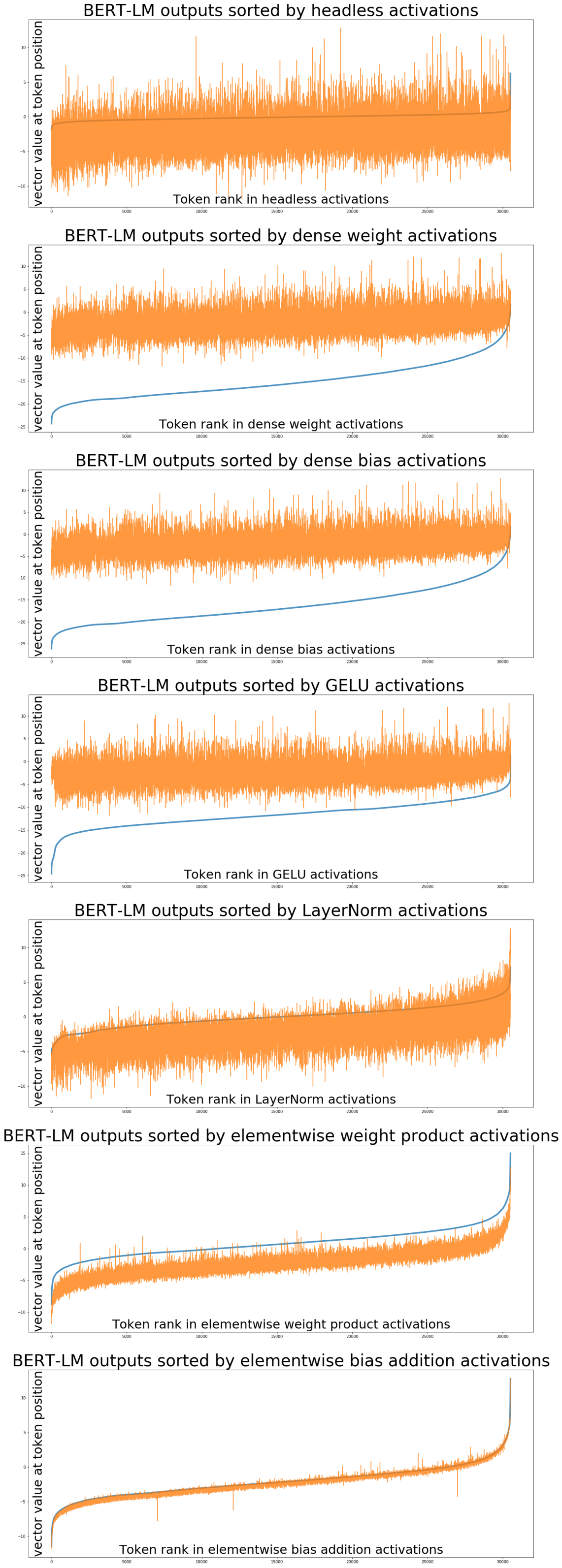

Fig. 4: BERT with LM head outputs sorted by inner products of token embeddings with each of the LM head activations (from top to bottom: encoder stack, dense weight, dense bias, GELU, LayerNorm, elementwise multiplication ($\otimes$) and addition ($\oplus$)). The plots visually align after the LayerNorm layer, and it coincides with reconstruction quality increase that we observe.

	trf-out	dense-w	dense-b	gelu	lrnorm-n	lrnorm-w	lrnorm-b	lmhead-out
trf-out	1	0.1264	0.1239	0.06393	0.0886	0.2204	0.257	0.2567
dense-w	0.1264	1	0.9994	0.7123	0.4282	0.13	0.3259	0.3674
dense-b	0.1239	0.9994	1	0.7197	0.4283	0.1217	0.3184	0.36
gelu	0.06393	0.7123	0.7197	1	0.07817	-0.04211	0.1888	0.1989
lrnorm-norm	0.0886	0.4282	0.4283	0.07817	1	0.6952	0.66	0.6815
lrnorm-w	0.2204	0.13	0.1217	-0.04211	0.6952	1	0.9462	0.9361
lrnorm-b	0.257	0.3259	0.3184	0.1888	0.66	0.9462	1	0.9962
lmhead-out	0.2567	0.3674	0.36	0.1989	0.6815	0.9361	0.9962	1

TABLE V: Pairwise Spearman's rank-order correlations (transformer stack output, dense weight output, dense bias output, gelu output, layernorm normalization output, layernorm elementwise multiplication output, layernorm elementwise bias addition output, LM head output).

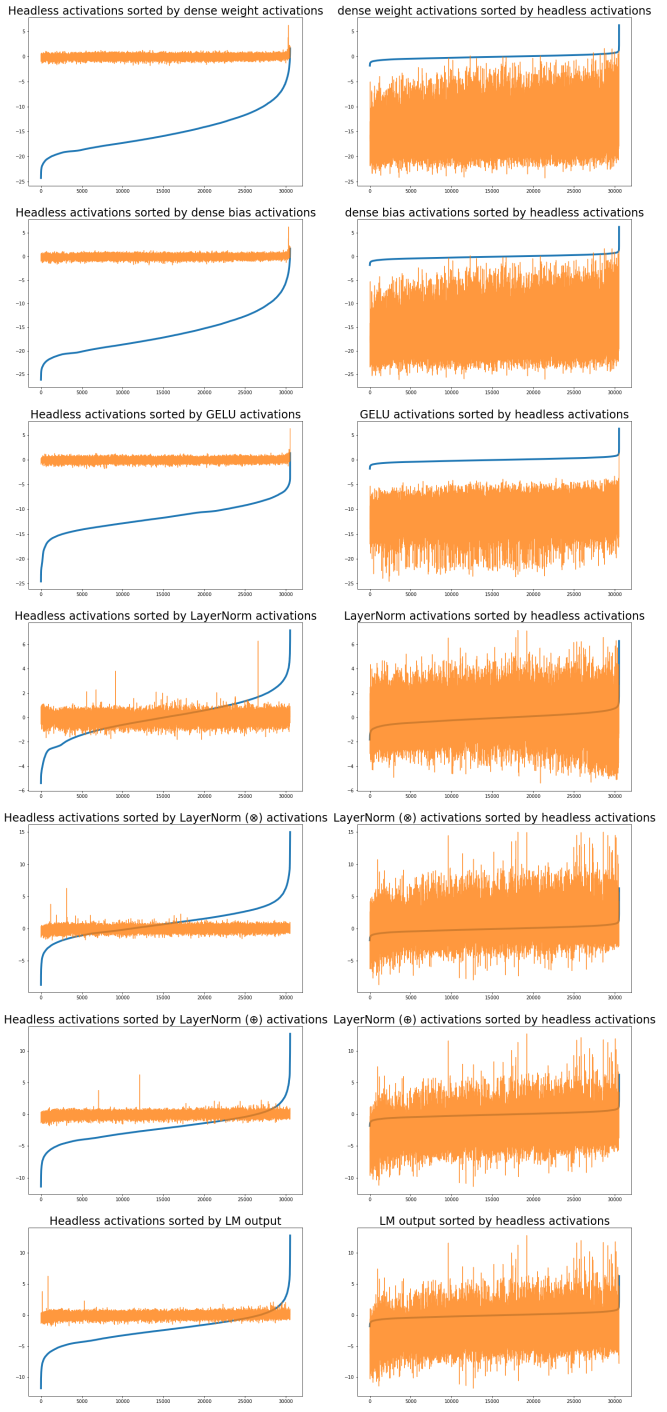

Fig. 5: LM head components compared to transformer output.

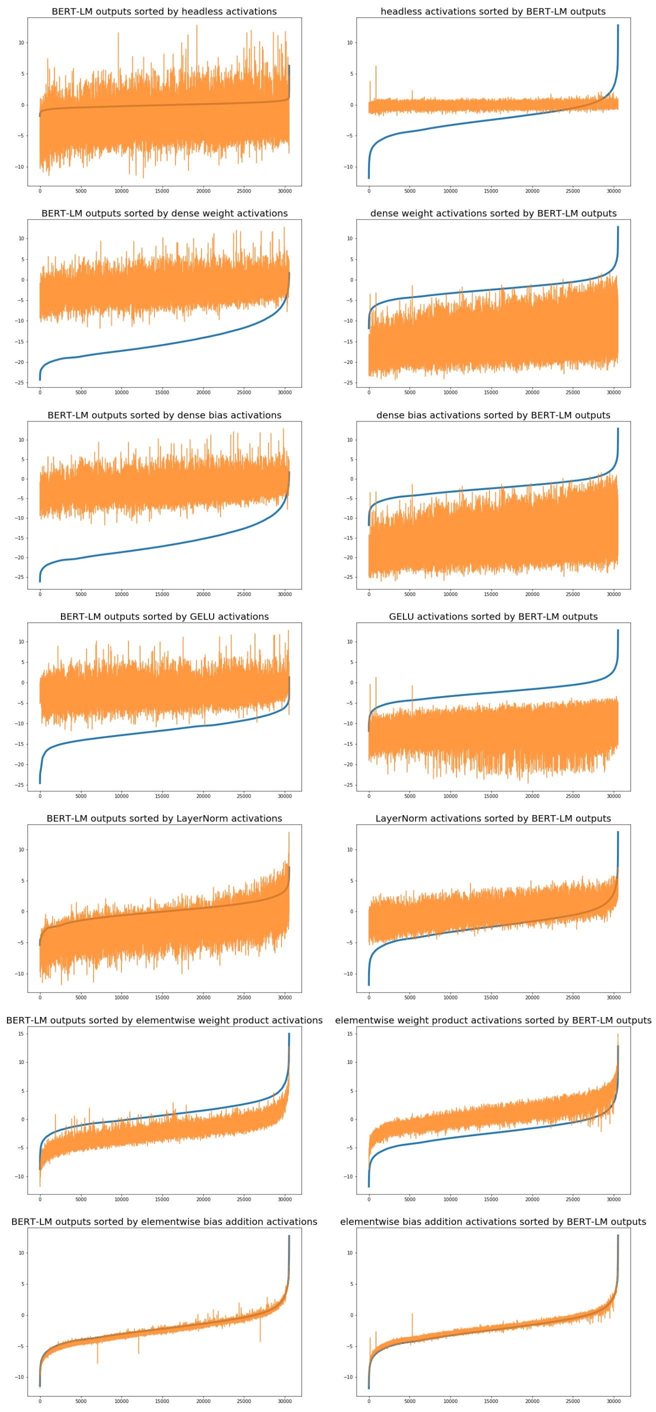

Fig. 6: LM head components compared to LM head output.

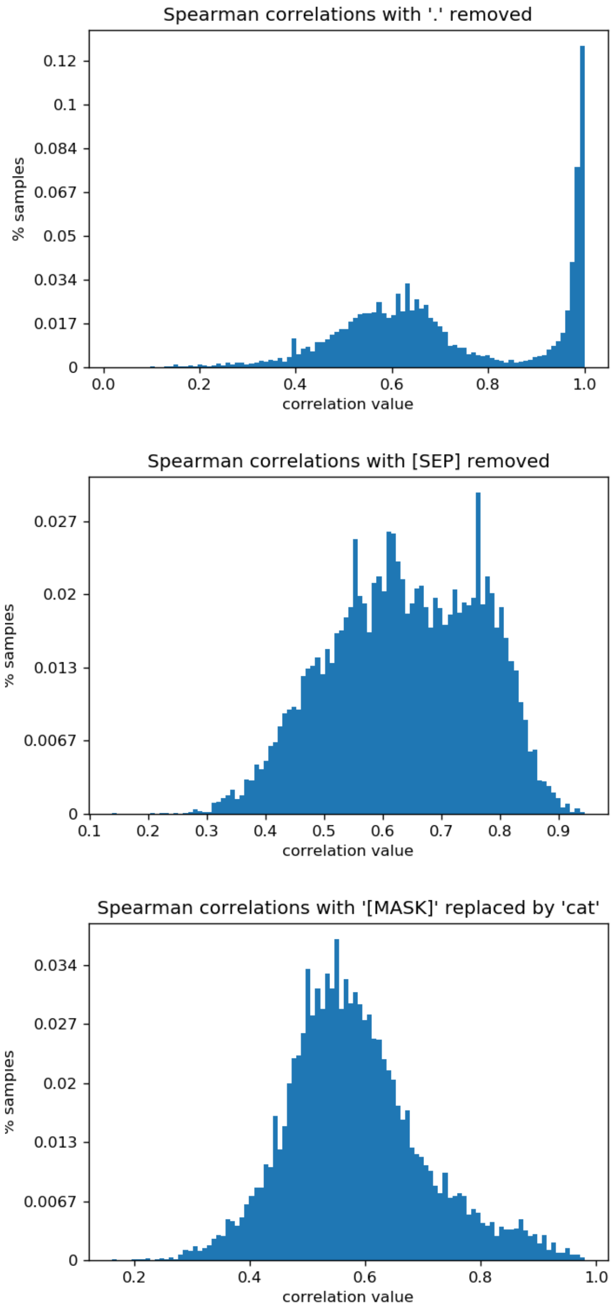

Fig. 7: Histogram of Spearman's rank-order correlations between logits of the original sample and logits of a perturbed sample, collected over entire ConceptNet dataset used in LAMA. In the top figure, we remove the '.' token ending the sample sentence or sentences; in the middle one – we remove the [SEP] token ending the token sequence serving as input to BERT; in the bottom one – we replace the [MASK] token with the 'cat' one. The decrease in reconstruction quality between perturbation strategies is reflected in the correlations decrease as the bell curve shifts to the left, if we compare the plots from top to bottom.

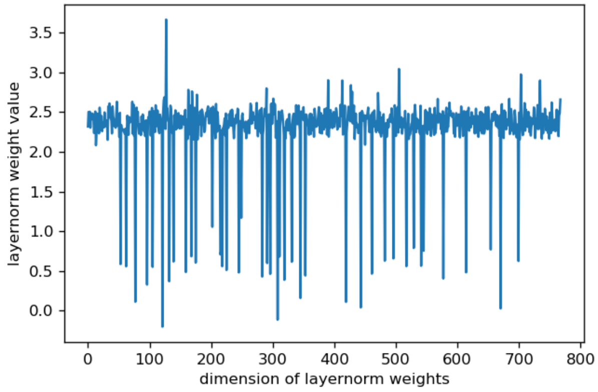

Fig. 8: LayerNorm weights in BERT base uncased.

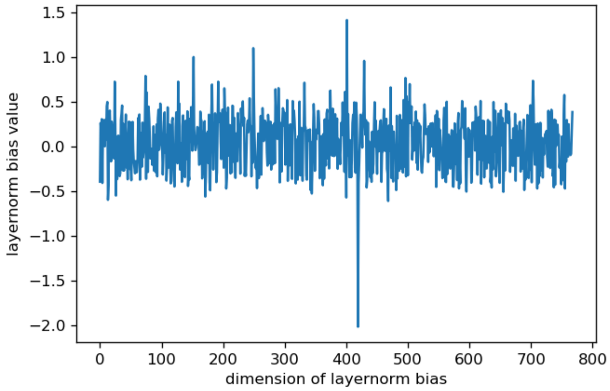

Fig. 9: LayerNorm bias in BERT base uncased.