10 Appendix

10.1 Candidate outlier dimensions of the models.

For each of the BERT-like models we experimented with we present the outlier candidate weights, detected as described in section 3.

Model component	Outliers
output.dense.weight	275, 276, 444
attention.output.dense.bias	193
attention.output.LayerNorm.weight	275, 276, 444
attention.output.LayerNorm.bias	276, 444
output.LayerNorm.weight	121, 262, 444, 276
output.LayerNorm.bias	276, 444

Table 7: BERT-small outlier dimension candidates across model components.

Model component	Outliers
attention.output.dense.bias	92, 400, 476, 17
output.dense.weight	400
output.dense.bias	400
attention.output.LayerNorm.weight	17, 400, 430
attention.output.LayerNorm.bias	192, 400
output.LayerNorm.weight	11, 193, 393, 427, 400
output.LayerNorm.bias	400, 427

Table 8: BERT-medium outlier dimension candidates across model components.

Model component	Outliers
output.dense.weight	308, 381
output.dense.bias	308
attention.output.dense.bias	308
attention.output.LayerNorm.weight	308, 381
attention.output.LayerNorm.bias	145, 308, 381
output.LayerNorm.weight	92, 145, 308, 381, 225
output.LayerNorm.bias	308, 381

Table 9: BERT-base outlier dimension candidates across model components.

10.2 Scaling factor and bias statistics for BERT-base.

For the BERT-base configuration, we present the detailed statistics on per-layer scaling factors and biases of the output LayerNorm (see Table 13). We report per-layer means, standard deviations and counts of the weights falling out of the three sigma range. We also show the values of the outlier weights (`308` and `381`) along with their ranks, where the ranks are computed for the corresponding sorted arrays of weight magnitudes. Note that the outlier weights consistently appear to be *among* the top largest or top smallest LayerNorm weights throughout the model, but are not necessarily the top-1 largest/smallest values.

10.3 Sample language model outputs after disabling outlier LayerNorm weights.

For RoBERTa and BERT, we randomly sample a set of sentences from Wikipedia and BookCorpus, mask multiple input tokens, and use the models for token prediction. We compare the baseline (full) models with the models where select LayerNorm weights are zeroed out across all of the Transformer layers. In particular, we compare the setups where the outlier dimensions (two per model) are disabled as opposed to random dimensions (two per model).

Model component	Outliers
attention.output.dense.bias	757, 327
output.dense.weight	757, 159
output.dense.bias	159, 757
attention.output.LayerNorm.weight	159, 757, 327
attention.output.LayerNorm.bias	159, 757, 327
output.LayerNorm.weight	159, 757, 327
output.LayerNorm.bias	159, 757, 327

Table 10: Multilingual BERT (mBERT) outlier dimension candidates across model components.

Model component	Outliers
output.dense.weight	588
output.dense.bias	588, 494
attention.output.dense.bias	588
attention.output.LayerNorm.bias	77, 217, 453, 551, 588, 496, 731, 494
output.LayerNorm.bias	77, 453, 551, 588, 217, 240, 496, 61, 494

Table 11: Base RoBERTa outlier dimensions across model components.

Model component	Outliers
attention.output.dense.bias	466, 18
output.dense.bias	466, 750, 18, 933
attention.output.LayerNorm.weight	234, 466, 933
attention.output.LayerNorm.bias	9, 71, 136, 234, 327, 706, 466, 474, 929, 933, 18, 143
output.LayerNorm.weight	80, 136, 232, 234, 331, 466, 639, 665, 702, 724, 750, 763, 968, 315, 428, 933, 18, 506, 314
output.LayerNorm.bias	136, 466, 706, 327, 9, 929, 18, 143, 933

Table 12: BERT-large outlier dimension candidates across model components.

	Scaling factors				Biases			
Transf. layer	mean / std	#$> 3\sigma$	308 value / rank	381 value / rank	mean / std	#$> 3\sigma$	308 value / rank	381 value/rank
1	0.756 / 0.056	12	0.343 / 764	0.404 / 762	-0.037 / 0.099	6	-1.325 / 0	0.144 / 78
2	0.870 / 0.069	24	0.400 / 765	0.374 / 766	-0.034 / 0.086	8	-0.678 / 0	0.277 / 5
3	0.851 / 0.052	16	0.408 / 767	0.549 / 765	-0.031 / 0.075	4	-0.070 / 298	0.118 / 103
4	0.811 / 0.044	11	0.562 / 764	0.388 / 767	-0.033 / 0.052	7	0.075 / 174	0.114 / 50
5	0.840 / 0.045	8	0.615 / 763	0.360 / 767	-0.031/ 0.051	8	0.200 / 3	-0.083 / 113
6	0.832 / 0.037	7	0.692 / 763	0.411 / 767	-0.032 / 0.060	6	0.403 / 0	-0.394 / 1
7	0.834 / 0.037	4	0.752 / 752	0.375 / 767	-0.033 / 0.063	5	0.785 / 0	-0.337 / 1
8	0.810 / 0.030	4	1.163 / 0	0.335 / 767	-0.033 / 0.065	2	0.959 / 0	0.304 / 1
9	0.831 / 0.042	6	1.618 / 0	0.262 / 767	-0.035 / 0.062	2	0.129 / 38	0.695/0
10	0.801 / 0.060	7	1.437 / 0	0.254 / 764	-0.032 / 0.057	9	-0.415 / 2	0.258/4
11	0.817 / 0.062	9	1.671 / 0	0.185 / 765	-0.040 / 0.068	5	-0.667 / 1	1.234/0
12	0.633 / 0.027	13	0.273 / 767	0.536 / 758	-0.019 / 0.050	5	0.225 / 0	-0.021/531

Table 13: The statistics of output LayerNorm weights (scaling factors and biases) for all of the Transformer layers of BERT-base.

Input	Ghostbusters was [released] on June 8 , [1984] , to critical [acclaim] and became a cultural phenomenon . It was well [received] for its deft blend of comedy, [action] , and horror , and Murray ' s performance was [repeatedly] singled out for praise .	a filmy coating of [dust] and pebbles had settled onto the block , and [sami] 's hand instinctively jerked forward to swipe the [scratchy] debris off his cheek , then pulled [up] short against the biting [metal] cuffs .	According to the RIAA, the Beatles are the best-[selling] music artists in the United States, with 178 [million] certified units. They have had more number-[one] albums on the [British] charts and sold [more] singles in the UK than any other act.
RoBERTa	Ghostbusters was [released] on June 8 , [1986] , to critical [acclaim] and became a cultural phenomenon . It was well [received] for its deft blend of comedy, [action] , and horror , and Murray ' s performance was [often] singled out for praise .	a filmy coating of [dirt] and pebbles had settled onto the block , and [Sami] 's hand instinctively jerked forward to swipe the [crusty] debris off his cheek , then pulled [up] short against the biting [leather] cuffs .	According to the RIAA, the Beatles are the best-[selling] music artists in the United States, with 178 [million] certified units. They have had more number-[one] albums on the [US] charts and sold [more] singles in the UK than any other act.
Random	Ghostbusters was [released] on June 8 , [1986] , to critical [acclaim] and became a cultural phenomenon . It was well [received] for its deft blend of comedy, [action] , and horror , and Murray ' s performance was [particularly] singled out for praise.	a filmy coating of [dirt] and pebbles had settled onto the block , and [Tsui] 's hand instinctively jerked forward to swipe the [crusty] debris off his cheek , then pulled [up] short against the biting [leather] cuffs .	According to the RIAA, the Beatles are the best-[selling] music artists in the United States, with 178 [million] certified units. They have had more number-[one] albums on the [US] charts and sold [more] singles in the UK than any other act.
Outliers	{ lock was [never] on June 8 , [</s>] , to rely [,] and . It was well [known] for its acker of comedy , [dinner], and horror , and Murray ' s was [ever] , </s> </s>)	a Fre) covering of [humor] and celebcele had </s> </s> </s> </s> , and [</s>i] 's </s> </s> </s> </s> </s> </s> (@ the [brainy] during (@ end) , Then pulled [*] isk ss the wearing [of] cuffs </s>	2017 </s> the RIAA, the Beatles are the [l] music files in the United States, with 178 [Canadian] Certified ols </s> They have had é yl-[million] Deaths on the [Chart] charts and Died [are] Hearts in</s> UK . </s></s></s></s>

Table 14: RoBERTa's masked language model predictions for randomly sampled input sequences. Input masked tokens (blue) are given in brackets. Correctly predicted tokens are shown in green, incorrect but plausible predictions are shown in brown. 48 weights have been modified in total for the *Random* and *Outliers* setups.

Input	he didnt [really] have a plan and he wasnt sure he [could] go through [with] anything , but the [feeling] of doing something was lifting his [spirits] .	ice is water frozen into a [solid] state . [depending] on the presence of impurities such as particles of soil or [bubbles] of air , it can appear [transparent] or a more or less [opaque] bluish - white color .	but the [sound] of the river babbling by the yard and the ducks splashing on the [pond] seemed to be [working] a cure for her [melancholy] .
BERT	he didnt [even] have a plan and he wasnt sure he [could] go through [with] anything , but the [thought] of doing something was lifting his [spirits] .	ice is water frozen into a [frozen] state . [depending] on the presence of impurities such as particles of soil or [particles] of air , it can appear [white] or a more or less [uniform] bluish - white color .	but the [sound] of the river babbling by the yard and the ducks splashing on the [water] seemed to be [providing] a cure for her [fears] .
Random	he didnt [even] have a plan and he wasnt sure he [could] go through [with] anything , but the [thought] of doing something was lifting his [spirits] .	ice is water frozen into a [liquid] state . [depending] on the presence of impurities such as particles of soil or [particles] of air , it can appear [white] or a more or less [uniform] bluish - white color .	but the [sounds] of the river babbling by the yard and the ducks splashing on the [water] seemed to be [just] a cure for her [fears] .
Outliers	he didny [wee] have a plan and he wasnt sure he [would] go through [it] anything , but the [actual] of doing something was lifting his [shoulders] .	that is water turned into a [yu] state . [based] on the presence of impurities such as particles of soil or [breath] of air , it can appear [white] or a more or more [commoning] ing - white color .	but the [sound] of the child babble by the yard and the ducks splashing on the [windows] all to be [in] a replacement for her [ness] .

Table 15: BERT's masked language model predictions for randomly sampled input sequences. Input masked tokens (blue) are given in brackets. Correctly predicted tokens are shown in green, incorrect but plausible predictions are shown in brown. 48 weights have been modified in total for the *Random* and *Outliers* setups.