A. Model Performance

We include a uniform activation baseline to contextualize the attention mechanism's utility. Table 3 notes the gap between the performance of uniform and softmax attention in the BiLSTM is never higher than 1-2%. This distinction aligns with the results of Wiegreffe and Pinter [27] and Vashishth et al. [59], who argue claims of attention-based interpretability are stronger in situations in which models need the attention module to solve the underlying task. Of course, it is difficult to prove a causal effect in a deep network: the BiLSTM may solve the task differently depending on whether or not the attention mechanism is ablated, meaning it is still possible to make claims of interpretability whether or not softmax attention leads to higher task performance. We again emphasize that we do not wish to take a side in the "attention explanation" debate, but this point further stresses the difficulty of proving anything with the *agreement as evaluation* paradigm.

Table 3. Test set accuracy using uniform and softmax activations in the attention mechanisms. A uniform activation of the attention weights makes the attention weights meaningless and, therefore, the drop in evaluation performance caused by using uniform attention weights gives an indication of the utility of the use attention layer(s) for each task.

	BiLSTM		DistillBERT	
	Uniform	Softmax	Uniform	Softmax
MNLI	.659 ± .001	.667 ± .004	.599 ± .002	.779 ± .002
Quora	.829 ± .001	.830 ± .001	.832 ± .001	.888 ± .001
SNLI	.804 ± .004	.807 ± .002	.770 ± .005	.871 ± .001
IMDb	.874 ± .011	.872 ± .014	.879 ± .003	.890 ± .005
SST-2	.823 ± .008	.826 ± .011	.823 ± .004	.842 ± .003