We thank the reviewers for their perceptive and useful comments. Subsequent to our submission, we expanded our experiments to include more encoder-only tasks and more seq2seq generative task like summarization. As reported in Tab. 1 and Tab. 2, we found that **BIGBIRD achieves new state-of-the-art (SoTA) for summarization task**. Also, we **obtained TriviaQA results to establish a new SoTA with 84.50 F1 on full and 92.39 F1** on verified subset. Furthermore for classification BIGBIRD achieves better performance than BERT. We will include these expanded results and full details of the experimental setup/hyper-parameters in the final version of the paper with the extra page.

	Model	Arxiv			PubMed			BigPatent		
		R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
Prior Art	Attn-Seq2Seq	29.30	6.00	25.56	31.55	8.52	27.38	28.74	7.87	24.66
	Pntr-Gen-Seq2Seq	32.06	9.04	25.16	35.86	10.22	29.69	33.14	11.63	28.55
	Long-Doc-Seq2Seq	35.80	11.05	31.80	38.93	15.37	35.21	-	-	-
	Sent-CLF	34.01	8.71	30.41	45.01	19.91	41.16	36.20	10.99	31.83
	Sent-PTR	42.32	15.63	38.06	43.30	17.92	39.47	34.21	10.78	30.07
	Extr-Abst-TLM	41.62	14.69	38.03	42.13	16.27	39.21	38.65	12.31	34.09
	Dancer	42.70	16.54	38.44	44.09	17.69	40.27	-	-	-
Base	Transformer	28.52	6.70	25.58	31.71	8.32	29.42	39.66	20.94	31.20
	+ RoBERTa	31.98	8.13	29.53	35.77	13.85	33.32	41.11	22.10	32.58
	+ Pegasus	34.81	10.16	30.14	39.98	15.15	35.89	43.55	20.43	31.80
	BIGBIRD-RoBERTa	41.22	16.43	36.96	43.70	19.32	39.99	55.69	37.27	45.56
Large	Pegasus (Reported)	44.21	16.95	38.83	45.97	20.15	41.34	52.29	33.08	41.75
	Pegasus (Re-eval)	43.85	16.83	39.17	44.53	19.30	40.70	52.25	33.04	41.80
	BIGBIRD-Pegasus	**46.63**	**19.02**	**41.77**	**46.32**	**20.65**	**42.33**	**60.64**	**42.46**	**50.01**

Table 1: Summarization ROUGE score for long documents.

Model	IMDb	Yelp-5	Arxiv	Patents	Hyperpartisan
SoTA	97.4	73.28	87.96	69.01	90.6
RoBERTa	95.0 ± 0.2	71.75	87.42	67.07	87.8 ± 0.8
BIGBIRD	95.2 ± 0.2	72.16	**92.31**	69.30	$\mathbf{92.2 \pm 1.7}$

Table 2: Classification results. We report the F1 micro-averaged score for all datasets.

Next we will answer the specific questions asked by each reviewers.

R2, Modeling relationship between tokens in different paragraphs: We agree that window attention models "locality of reference", but global attention captures long distance relationships. BIGBIRD also uses random attention which is motivated from the ability of random graphs to capture properties of fully connected graphs, hence adding another way to capture long distance relationships. Moreover, our theoretical analysis shows that BIGBIRD is able to capture all sequence to sequence functions, including the ones that have long range dependency, while our empirical results back this claim by outperforming baselines.

R2, Inference Time: We compared BIGBIRD and BERT on sequences of length 512 and 1024 and found the inference time to be comparable. BERT uses the full attention mechanism and thus goes out of memory for sequence with more than 1K tokens. We will include this in the paper.

R3, Experiments on shorter text: Results from experiments on shorter text have been reported in table 15, section E.4 in appendix. The table compares performance of BIGBIRD on 8 different General Language Understanding Evaluation (GLUE) benchmark tasks. We see that BIGBIRD performs competitively even on smaller input sequences.

R4, Related work: While window attention models have been proposed before, prior models were based on heuristics and were not as versatile and robust as the original transformer. In particular, the same architecture did not attain SoTA on multiple standard benchmarks nor handle both encoding and decoding. Moreover, these approximations did not come with any theoretical guarantees. We both extend the theoretical understanding of sparse models and provide BIGBIRD-attention architecture that achieves SoTA for multiple applications.

R4, Quadratic to Linear: We made an asymptotic statement assuming window size is constant as sequence length grows. In particular, for N tokens the total number of attention in BIGBIRD is upper bounded $N(2b + w + r)$ instead of N^2 in BERT and transformers. Here b, w, r are the size of global attention, local attention and random attention per query respectively. These sizes are kept constant for all the experiments leading to attention being linear in the number of token. We have reported results when $N = 4096$, but have conducted experiments with $N > 16,000$ tokens, where $N \gg 2b + w + r$ and asymptotic behaviour kicks in. We will add these details and clarify further in the main text.