C Rebuttal Results Summary

Table 4: In this table, we summarize perplexity on C4 with different sequence lengths.

	[256 - 512]	[512- 1024]	[1024- 2048]
OPT-13B	8.7968	9.1017	9.3005
OPT-13B + SCISSORHANDS	8.7972	9.1011	9.3009
OPT-13B + Local Window	81.8297	29.3823	15.5883

Table 5: This table summarizes the raw accuracy number on tested datasets.

Models	Hellaswag	MathQA	PIQA	Winogrande
OPT-6B	0.702	0.199	0.775	0.606
OPT-6B + Scissorhands	0.694	0.195	0.773	0.606
OPT-13B	0.726	0.206	0.780	0.613
OPT-13B + Scissorhands	0.720	0.209	0.780	0.614
OPT-30B	0.752	0.211	0.784	0.649
OPT-30B + Scissorhands	0.756	0.209	0.784	0.650

Table 6: This table summarizes the accuracy of three additional datasets.

Models	OpenBookQA	RTE	COPA
OPT-13B	0.382	0.527	0.83
OPT-13B + Scissorhands	0.382	0.527	0.81

Table 7: This table provides generated examples using OPT-13B with full cache and SCISSORHANDSat different compression ratio.

Prompt	Hi. Does Asus ever plan to update the intel and nvidia display drivers on the UL30VT laptop? The only downloadable drivers from the Asus website are at least 3 years behind and new features for both graphics chips have been released since then. I sent a message to Asus but no response was given. Intel and Nvidia will not allow the user to download the new drivers directly. problem is the drivers won't install. states something about your notebook needs custom drivers from the manufacturer due to customized this that or the other.
OPT-13B	I have the same problem. I have a laptop with an Intel HD Graphics 4000 and an Nvidia GeForce GT 740M. I can't install the drivers from the Asus website. I have to use the drivers from the Nvidia website.
3x Compression	I have the same problem. I have a laptop with an Intel HD Graphics 4000 and an Nvidia GeForce GT 740M. I can't install the drivers from the Asus website. I have to use the drivers from the Nvidia website.
6x Compression	I have the same problem. I have a custom laptop and I can't install the drivers.

Table 8: This table provides generation throughput measured in tokens per second on OPT-6B and OPT-13B. We use a prompt length of 512 and a generation length of 1024.

Model	OPT-6B	OPT-13B
FlexGen	40.1	28.4
FlexGen + SCISSORHANDS	42.2	33.9