0.1 Full Results

In this section, we list the full experiment results for each individual dataset used.

Wikitext (GPT2)	Perplexity	Change in Perplexity (%)	Runtime (s)	Change in Runtime (%)	Energy (KwH)	Change in Energy (%)
Base	31.02443		191		0.02078	
Base8b	31.10938	0.274%	581	204.188%	0.009257	-55.450%
Base4b	32.28125	4.051%	200	4.712%	0.010311	-50.379%
Distil	48.9968	57.930%	119	-37.696%	0.012353	-40.550%
AH90	49.58415	59.823%	182	-4.712%	0.019847	-4.488%
AH80	59.14925	90.654%	176	-7.853%	0.019091	-8.127%
Distil+90	54.1907	74.671%	115	-39.791%	0.011846	-42.994%
Distil+80	67.68159	118.156%	111	-41.885%	0.011351	-45.374%
Distil8b	49.21875	58.645%	283	48.168%	0.004929	-76.281%
Distil4b	50.40625	62.473%	113	-40.838%	0.006023	-71.013%
AH90+8b	49.21875	58.645%	581	204.188%	0.008872	-57.303%
AH90+4b	50.59375	63.077%	194	1.571%	0.010633	-48.830%
AH80+8b	59.5	91.784%	569	197.906%	0.008603	-58.599%
AH80+4b	61.15625	97.123%	196	2.618%	0.010011	-51.821%
Distil+90+8b	54.0625	74.258%	271	41.885%	0.004691	-77.424%
Distil+90+4b	56.125	80.906%	111	-41.885%	0.00573	-72.424%
Distil+80+8b	68.1875	119.786%	262	37.173%	0.004421	-78.724%
Distil+80+4b	69	122.405%	109	-42.932%	0.005442	-73.812%
OPT-125m	36.09		5291		0.271119	
Pruned	45.32475	25.588%	4146	-21.641%	0.244033	-9.991%

Table 1: Results of Perplexity tests performed on the GPT2 model using the Wikitext-2 dataset.

C4 (GPT2)	Perplexity	Change in Perplexity (%)	Runtime (s)	Change in Runtime (%)	Energy (KwH)	Change in Energy (%)
Base	32.71996		162		0.01767	
Base8b	32.84375	0.378%	493	204.321%	0.007768	-56.037%
Base4b	34.3125	4.867%	170	4.938%	0.008847	-49.932%
Distil	44.18049	35.026%	101	-37.654%	0.010559	-40.243%
AH90	38.62528	18.048%	156	-3.704%	0.016963	-4.005%
AH80	43.57146	33.165%	151	-6.790%	0.016323	-7.622%
Distil+90	50.55785	54.517%	99	-38.889%	0.010159	-42.511%
Distil+80	57.65889	76.219%	95	-41.358%	0.009698	-45.119%
Distil8b	44.3125	35.430%	237	46.296%	0.004154	-76.492%
Distil4b	46.34375	41.638%	96	-40.741%	0.005155	-70.827%
AH90+8b	38.78125	18.525%	494	204.938%	0.007628	-56.831%
AH90+4b	40.96875	25.210%	169	4.321%	0.008867	-49.817%
AH80+8b	43.6875	33.519%	489	201.852%	0.00741	-58.063%
AH80+4b	45.53125	39.154%	166	2.469%	0.008221	-53.474%
Distil+90+8b	50.78125	55.200%	233	43.827%	0.004018	-77.263%
Distil+90+4b	53.21875	62.649%	95	-41.358%	0.005137	-70.929%
Distil+80+8b	57.90625	76.975%	222	37.037%	0.003778	-78.622%
Distil+80+4b	60.4375	84.711%	94	-41.975%	0.004565	-74.165%
OPT-125m	28.64		4504		0.235593	
Pruned	45.82689	60.010%	3547	-21.248%	0.185149	-21.411%

Table 2: Results of Perplexity tests performed on the GPT2 model using the C4 dataset.

Lambada (GPT2)	Perplexity	Change in Perplexity (%)	Runtime (s)	Change in Runtime (%)	Energy (KwH)	Change in Energy (%)
Base	39.13642		284		0.031606	
Base8b	39.25	0.290%	860	202.817%	0.013476	-57.361%
Base4b	40.09375	2.446%	298	4.930%	0.015704	-50.313%
Distil	55.05059	40.663%	175	-38.380%	0.018926	-40.119%
AH90	48.59665	24.172%	273	-3.873%	0.029864	-5.511%
AH80	54.00827	38.000%	266	-6.338%	0.02896	-8.372%
Distil+90	63.34478	61.856%	174	-38.732%	0.0186	-41.152%
Distil+80	70.19104	79.350%	168	-40.845%	0.017905	-43.349%
Distil8b	55.25	41.173%	417	46.831%	0.007271	-76.996%
Distil4b	57	45.644%	168	-40.845%	0.009239	-70.767%
AH90+8b	48.5625	24.085%	841	196.127%	0.013014	-58.825%
AH90+4b	49.53125	26.560%	295	3.873%	0.016065	-49.170%
AH80+8b	54.1875	38.458%	849	198.944%	0.013042	-58.737%
AH80+4b	55.46875	41.732%	290	2.113%	0.015122	-52.154%
Distil+90+8b	63.84375	63.131%	411	44.718%	0.007155	-77.364%
Distil+90+4b	65.875	68.321%	165	-41.901%	0.009189	-70.928%
Distil+80+8b	70.125	79.181%	405	42.606%	0.006835	-78.374%
Distil+80+4b	72.0625	84.132%	166	-41.549%	0.00888	-71.903%
OPT-125m	39.41		6697		0.597286	
Pruned	57.3837	45.607%	6257	-6.570%	0.394584	-33.937%

Table 3: Results of Perplexity tests performed on the GPT2 model using the Lambada dataset.

Wikitext (L)	Perplexity	Change in Perplexity (%)	Runtime (s)	Change in Runtime (%)	Energy (KwH)	Change in Energy (%)
Base	24.9575		964		0.110359	
Base8b	24.95313	-0.018%	1760	82.573%	0.034199	-69.011%
Base4b	25.34375	1.548%	654	-32.158%	0.054515	-50.602%
Distil	39.89427	59.849%	509	-47.199%	0.057927	-47.511%
AH90	31.77251	27.306%	873	-9.440%	0.099753	-9.611%
AH80	36.6982	47.043%	825	-14.419%	0.093654	-15.137%
Distil+90	40.40304	61.887%	507	-47.407%	0.057832	-47.597%
Distil+80	40.35398	61.691%	502	-47.925%	0.056842	-48.494%
Distil8b	39.9375	60.022%	863	-10.477%	0.020297	-81.608%
Distil4b	40.5	62.276%	340	-64.730%	0.02851	-74.166%
AH90+8b	31.78125	27.341%	1667	72.925%	0.028897	-73.816%
AH90+4b	32.65625	30.847%	636	-34.025%	0.048672	-55.897%
AH80+8b	36.71875	47.125%	1752	81.743%	0.030657	-72.221%
AH80+4b	37.3125	49.504%	624	-35.270%	0.046007	-58.311%
Distil+90+8b	40.5	62.276%	852	-11.618%	0.019928	-81.942%
Distil+90+4b	41.21875	65.156%	336	-65.145%	0.027576	-75.013%
Distil+80+8b	40.40625	61.900%	862	-10.581%	0.019497	-82.333%
Distil+80+4b	40.90625	63.904%	344	-64.315%	0.027489	-75.091%

Table 4: Results of Perplexity tests performed on the GPT2-L model using the Wikitext-2 dataset.

C4 (L)	Perplexity	Change in Perplexity (%)	Runtime (s)	Change in Runtime (%)	Energy (KwH)	Change in Energy (%)
Base	24.57314		816		0.093666	
Base8b	24.5625	-0.043%	1508	84.804%	0.028864	-69.184%
Base4b	25	1.737%	555	-31.985%	0.047175	-49.635%
Distil	34.95628	42.254%	438	-46.324%	0.049671	-46.970%
AH90	27.47602	11.813%	745	-8.701%	0.084925	-9.332%
AH80	30.82773	25.453%	705	-13.603%	0.079878	-14.721%
Distil+90	34.73307	41.346%	435	-46.691%	0.048666	-48.043%
Distil+80	34.93757	42.178%	429	-47.426%	0.048476	-48.246%
Distil8b	34.96875	42.305%	752	-7.843%	0.016856	-82.004%
Distil4b	35.65625	45.103%	295	-63.848%	0.024627	-73.708%
AH90+8b	27.5	11.911%	1505	84.436%	0.027229	-70.930%
AH90+4b	28.04688	14.136%	545	-33.211%	0.043746	-53.296%
AH80+8b	30.8125	25.391%	1509	84.926%	0.026654	-71.543%
AH80+4b	31.46875	28.062%	533	-34.681%	0.039153	-58.200%
Distil+90+8b	34.78125	41.542%	712	-12.745%	0.017061	-81.785%
Distil+90+4b	35.25	43.449%	295	-63.848%	0.023716	-74.681%
Distil+80+8b	34.96875	42.305%	746	-8.578%	0.015818	-83.112%
Distil+80+4b	35.46875	44.339%	294	-63.971%	0.023622	-74.781%

Table 5: Results of Perplexity tests performed on the GPT2-L model using the C4 dataset.

Lambada (L)	Perplexity	Change in Perplexity (%)	Runtime (s)	Change in Runtime (%)	Energy (KwH)	Change in Energy (%)
Base	31.25605		1428		0.164443	
Base8b	31.23438	-0.069%	2619	83.403%	0.050443	-69.325%
Base4b	31.59375	1.080%	978	-31.513%	0.078645	-52.175%
Distil	55.30352	76.937%	762	-46.639%	0.086838	-47.193%
AH90	36.2423	15.953%	1320	-7.563%	0.151548	-7.842%
AH80	40.73369	30.323%	1241	-13.095%	0.141597	-13.893%
Distil+90	54.91515	75.694%	763	-46.569%	0.087698	-46.670%
Distil+80	57.18518	82.957%	757	-46.989%	0.086765	-47.237%
Distil8b	55.25	76.766%	1214	-14.986%	0.028248	-82.822%
Distil4b	55.90625	78.865%	514	-64.006%	0.043085	-73.799%
AH90+8b	36.3125	16.178%	2624	83.754%	0.048167	-70.709%
AH90+4b	36.4375	16.577%	920	-35.574%	0.073128	-55.530%
AH80+8b	40.8125	30.575%	2492	74.510%	0.041714	-74.633%
AH80+4b	41.28125	32.074%	931	-34.804%	0.068955	-58.068%
Distil+90+8b	55.03125	76.066%	1157	-18.978%	0.028688	-82.554%
Distil+90+4b	55.90625	78.865%	512	-64.146%	0.041155	-74.973%
Distil+80+8b	57.21875	83.065%	1188	-16.807%	0.028682	-82.558%
Distil+80+4b	58.34375	86.664%	493	-65.476%	0.040774	-75.205%

Table 6: Results of Perplexity tests performed on the GPT2-L model using the Lambada dataset.

Note that the Lambada dataset is excluded from the set of experiments performed on GPT2-XL, as the length of the training samples in said dataset make training impossible on any publicly-available system due to memory requirements.

Wikitext (XL)	Perplexity	Change in Perplexity (%)	Runtime	Change in Runtime (%)	Energy	Change in Energy (%)
Base	15.91		785		0.075034	
8b	15.92	-0.051%	812	3.439%	0.020523	-72.648%
4b	17.28	8.611%	383	-51.210%	0.033382	-55.511%
AH90	19.36	21.684%	505	-35.669%	0.05858	-21.928%
AH80	22.05	38.592%	473	-39.745%	0.054423	-27.469%
AH90+8b	19.39	21.873%	825	5.096%	0.016769	-77.651%
AH90+4b	19.62	23.319%	354	-54.904%	0.029442	-60.762%
AH80+8b	22.06	38.655%	825	5.096%	0.015954	-78.738%
AH80+4b	22.58	41.923%	344	-56.178%	0.027568	-63.259%

Table 7: Results of Perplexity tests performed on the GPT2-XL model using the Wikitext-2 dataset.

C4 (XL)	Perplexity	Change in Perplexity (%)	Runtime	Change in Runtime (%)	Energy	Change in Energy (%)
Base	19.43		525		0.058789	
8b	19.42	-0.051%	707	34.667%	0.017417	-70.374%
4b	20.92	7.669%	328	-37.524%	0.028823	-50.972%
AH90	20.72	6.639%	432	-17.714%	0.049896	-15.127%
AH80	22.47	15.646%	405	-22.857%	0.0466	-20.733%
AH90+8b	20.72	6.639%	711	35.429%	0.014327	-75.631%
AH90+4b	21.09	8.543%	305	-41.905%	0.026117	-55.576%
AH80+8b	22.5	15.800%	719	36.952%	0.01363	-76.816%
AH80+4b	22.89	17.808%	294	-44.000%	0.023504	-60.019%

Table 8: Results of Perplexity tests performed on the GPT2-XL model using the C4 dataset.