Table 21: TwinBreak performance using the pruned model during the entire response generation process.

Model	HarmBench [39]		JBB [8]	ADV [75]	SREJECT [57]	Average Difference in Utility Benchmarks with Clean Model				
	Train	Val				Winogrande	RTE	ARC Challenge	OpenBookQA	HellaSwag
LLaMA 2 (7B) [42]	91%	95%	92%	96.35%	0.708	+0.6%	-2.7%	-1.9%	-0.5%	0.0%
LLaMA 3.1 (8B) [43]	96%	99%	95%	98.08%	0.801	-1.9%	-10.1%	-3.5%	-7.3%	-4.5%
Qwen 2.5 (7B) [26]	97%	97%	96%	98.85%	0.782	-4.6%	-7.7%	-1.3%	-1.0%	-1.5%
Gemma 2 (9B) [19]	89%	97%	92%	94.62%	0.719	-2.3%	-0.7%	-3.0%	-0.2%	+1.2%

Table 22: Performance of directional ablation [4] jailbreak.

Model	HarmBench [39]		JBB [8]	ADV [75]	SREJECT [57]	Difference in Utility Benchmarks with Clean Model				
	Train	Val				Winogrande	RTE	ARC Challenge	OpenBookQA	HellaSwag
LLaMA 2 (7B) [42]	85%	87%	90%	90.38%	0.605	-0.5%	0.0%	0.0%	+1.5%	+0.5%
LLaMA 3.1 (8B) [43]	94%	95%	92%	95.00%	0.798	+1.0%	+2.5%	+0.5%	+0.5%	0.0%
Qwen 2.5 (7B) [26]	91%	93%	91%	93.26%	0.798	-1.5%	0.0%	-1.5%	1.5%	0.5%
Gemma 2 (9B) [19]	90%	93%	91%	94.42%	0.771	-0.5%	-2.5%	-1.5%	0.0%	+0.5%

Table 23: Set difference [67] performance on LLaMA 2 over the best-performing configurations reported in [67].

Params		HarmBench [39]		JBB [8]	ADV [75]	SREJECT [57]	Difference in Utility Benchmarks with Clean Model				
p	q	Train	Val				Winogrande	RTE	ARC Challenge	OpenBookQA	HellaSwag
1	1	76%	84%	77%	80.19%	0.355	+1.0%	-12.5%	-1.5%	-5.5%	-1.5%
2	1	71%	71%	66%	71.34%	0.365	-2.5%	-7.5%	-1.0%	-0.5%	0.0%
3	2	85%	93%	86%	86.92%	0.401	-4.0%	-4.5%	-2.5%	-6.5%	-2.5%
4	2	76%	86%	79%	81.34%	0.403	-6.0%	-1.5%	-3.5%	-4.5%	0.0%
4	4	94%	98%	96%	95.38%	0.241	+2.5%	-9.5%	-9.0%	-8.0%	-5.5%
5	5	93%	96%	97%	95.00%	0.226	-4.0%	-1.5%	-6.0%	-10.0%	-5.5%
6	5	87%	94%	92%	94.03%	0.313	-4.0%	-6.5%	-3.0%	-7.0%	-4.5%
6	6	92%	95%	96%	95.76%	0.224	-3.0%	-5.5%	-5.5%	-10.0%	-7.5%
7	3	75%	81%	68%	77.5%	0.365	-3.0%	+3.0%	-3.0%	-0.5%	-2.0%
9	8	91%	96%	94%	93.07%	0.305	-7.5%	-5.5%	-5.5%	-8.0%	-3.5%

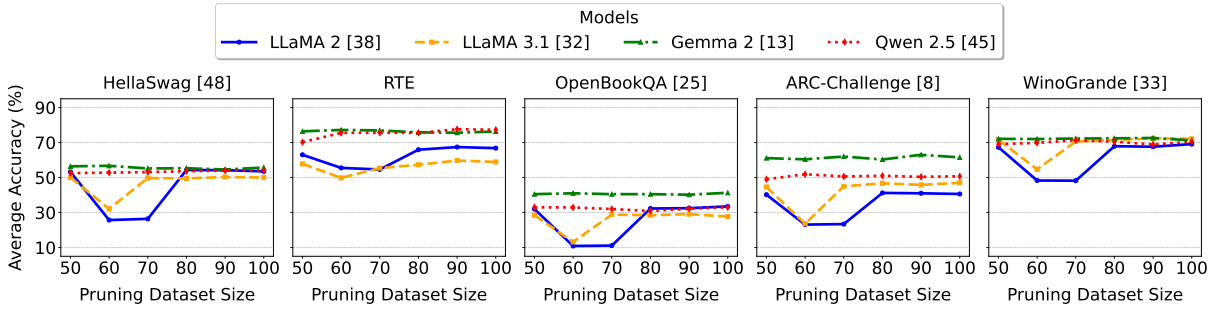

Figure 13: Evaluating the performance of utility benchmarks of various models when using different sizes of the pruning dataset.

Table 24: Results on varied the model sizes and families for TwinBreak on StrongREJECT [57] using the full pruning dataset.

Models	StrongREJECT	Difference in Utility Benchmarks with Clean Model				
		Winogrande	RTE	ARC Challenge	OpenBookQA	HellaSwag
LLaMA 2 7B	0.702	+0.6%	-2.7%	-1.9%	-0.5%	0.0%
LLaMA 2 13B	0.714	-5.3%	-5.2%	-3.9%	-4.4%	-0.4%
LLaMA 2 70B (8-bit)	0.674	-1.3%	-14.40%	-3.5%	-6.2%	-0.8%
LLaMA 3.1 8B	0.805	-1.9%	-10.10%	-3.5%	-7.3%	-4.5%
LLaMA 3.3 70B	0.762	-2.0%	0.0%	+1.59%	+0.20%	+0.69%
Gemma 2 2B	0.696	-4.5%	-14.8%	-4.9%	-2.1%	+0.3%
Gemma 2 9B	0.683	-2.3%	-0.7%	-3.0%	-0.2%	+1.2%
Gemma 2 27B	0.680	-3.9%	-0.1%	-0.2%	-1.6%	+0.3%
Gemma 3 1B	0.688	-4.8%	-14.49%	-9.2%	-5.2%	-2.19%
Qwen 2.5 3B	0.779	-8.4%	-21.6%	-5.0%	-5.1%	-4.7%
Qwen 2.5 7B	0.794	-4.6%	-7.7%	-1.3%	-1.0%	-1.5%
Qwen 2.5 14B	0.781	-1.8%	-2.9%	-5.4%	-0.6%	-3.1%
Qwen 2.5 32B	0.814	-3.3%	-2.8%	-0.7%	+1.6%	-1.2%
Qwen 2.5 72B (8-bit)	0.799	-5.5%	-5.0%	-2.3%	-1.8%	-1.1%
Mistral 7B	0.765	-4.0%	-7.3%	1.3%	-3.8%	-2.8%
DeepSeek 7B	0.773	-5.5%	-8.3%	2.29%	-4.19%	0.0%

Table 25: Results on varied the model sizes and families for Directional Ablation [4] on StrongREJECT [57].

Models	StrongREJECT	Difference in Utility Benchmarks with Clean Model				
		Winogrande	RTE	ARC Challenge	OpenBookQA	HellaSwag
LLaMA 2 7B	0.605	-0.5%	0.0%	0.0%	+1.5%	+0.5%
LLaMA 2 13B	0.188	-2.0%	0.0%	-1.0%	-1.5%	-1.5%
LLaMA 2 70B (8-bit)	0.345	-2.5%	+2.0%	-0.5%	-0.5%	-1.0%
LLaMA 3.1 8B	0.798	+1.0%	+2.5%	+0.5%	+0.5%	0.0%
LLaMA 3.3 70B	0.733	-1.5%	0.0%	-1.0%	-0.5%	0.0%
Gemma 2 2B	0.598	-7.0%	-7.5%	-13.0%	-10.5%	-6.5%
Gemma 2 9B	0.771	-0.5%	-2.5%	-1.5%	0.0%	+0.5%
Gemma 2 27B	0.000	-27.5%	-22.0%	-40.0%	-31.0%	-33.0%
Gemma 3 1B	0.000	-16.5%	-15.0%	-19.0%	-18.5%	-18.5%
Qwen 2.5 3B	0.516	-4.5%	-16.0%	-12.0%	-5.5%	-3.5%
Qwen 2.5 7B	0.798	-1.5%	0.0%	-1.5%	1.5%	0.5%
Qwen 2.5 14B	0.852	+3.0%	-4.0%	-1.5%	-2.0%	-2.5%
Qwen 2.5 32B	0.807	0.0%	-3.5%	-2.5%	-0.5%	-1.0%
Qwen 2.5 72B (8-bit)	0.713	+1.0%	-0.5%	-1.0%	-1.5%	-1.0%
Mistral 7B	0.756	-0.5%	+0.5%	-1.5%	-1.0%	-2.5%
DeepSeek 7B	0.778	+1.5%	-3.0%	+2.0%	+1.0%	+1.5%