

MMLU Acc

Lora –				7.0%	1.5%		0.5%	0.4%
Fine-tune –				0.3%	0.8%	0.6%	0.6%	1.7%
MEMIT –				0.3%	0.5%	1.8%	0.0%	0.4%
RMU –	7.0%	0.3%	0.3%		4.3%	1.3%	0.9%	0.2%
GPTQ –	1.5%	0.8%	0.5%	4.3%				
Wanda –		0.6%	1.8%	1.3%				
SparseGPT –	0.5%	0.6%	0.0%	0.9%				
AWQ –	0.4%	1.7%	0.4%	0.2%				
	Lora	Fine-tune	MEMIT	RMU	GPTQ	Wanda	SparseGPT	AWQ

WMDP Acc

Lora –				0.9%	2.7%		0.5%	0.4%
Fine-tune –				0.1%	1.6%	0.1%	0.2%	0.3%
MEMIT –				0.8%	0.0%	0.0%	1.6%	0.3%
RMU –	0.9%	0.1%	0.8%		10.6%	1.2%	1.0%	0.8%
GPTQ –	2.7%	1.6%	0.0%	10.6%				
Wanda –		0.1%	0.0%	1.2%				
SparseGPT –	0.5%	0.2%	1.6%	1.0%				
AWQ –	0.4%	0.3%	0.3%	0.8%				
	Lora	Fine-tune	MEMIT	RMU	GPTQ	Wanda	SparseGPT	AWQ

Edit Success

Lora –				0.0%	66.5%		18.1%	8.6%
Fine-tune –				0.5%	56.0%	1.2%	0.5%	1.5%
MEMIT –				0.8%	17.2%	24.7%	4.3%	16.4%
RMU –	0.0%	0.5%	0.8%		0.0%	0.0%	0.0%	0.4%
GPTQ –	66.5%	56.0%	17.2%	0.0%				
Wanda –		1.2%	24.7%	0.0%				
SparseGPT –	18.1%	0.5%	4.3%	0.0%				
AWQ –	8.6%	1.5%	16.4%	0.4%				
	Lora	Fine-tune	MEMIT	RMU	GPTQ	Wanda	SparseGPT	AWQ