

MMLU Acc

Lora –				0.5%	0.5%	1.5%	0.4%	
Fine-tune –				0.5%	0.6%	0.8%	1.7%	0.6%
MEMIT –				2.0%	0.0%	0.5%	0.4%	1.8%
RMU –	0.5%	0.5%	2.0%		4.1%	6.0%	5.6%	4.7%
SparseGPT –	0.5%	0.6%	0.0%	4.1%				
GPTQ –	1.5%	0.8%	0.5%	6.0%				
AWQ –	0.4%	1.7%	0.4%	5.6%				
Wanda –		0.6%	1.8%	4.7%				
	Lora	Fine-tune	MEMIT	RMU	SparseGPT	GPTQ	AWQ	Wanda

WMDP Acc

Lora –				3.2%	0.5%	2.7%	0.4%	
Fine-tune –				6.0%	0.2%	1.6%	0.3%	0.1%
MEMIT –				6.0%	1.6%	0.0%	0.3%	0.0%
RMU –	3.2%	6.0%	6.0%		9.9%	27.2%	2.2%	20.9%
SparseGPT –	0.5%	0.2%	1.6%	9.9%				
GPTQ –	2.7%	1.6%	0.0%	27.2%				
AWQ –	0.4%	0.3%	0.3%	2.2%				
Wanda –		0.1%	0.0%	20.9%				
	Lora	Fine-tune	MEMIT	RMU	SparseGPT	GPTQ	AWQ	Wanda

Edit Success

Lora –				0.0%	18.1%	66.5%	8.6%	
Fine-tune –				0.5%	0.5%	56.0%	1.5%	1.2%
MEMIT –				19.3%	4.3%	17.2%	16.4%	24.7%
RMU –	0.0%	0.5%	19.3%		0.0%	0.4%	1.3%	0.4%
SparseGPT –	18.1%	0.5%	4.3%	0.0%				
GPTQ –	66.5%	56.0%	17.2%	0.4%				
AWQ –	8.6%	1.5%	16.4%	1.3%				
Wanda –		1.2%	24.7%	0.4%				
	Lora	Fine-tune	MEMIT	RMU	SparseGPT	GPTQ	AWQ	Wanda