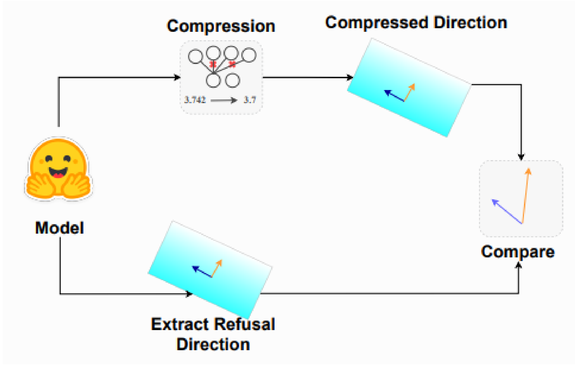
Computer Science and Engineering

When Compression Breaks Safety, Interpretability Reveals The Fix

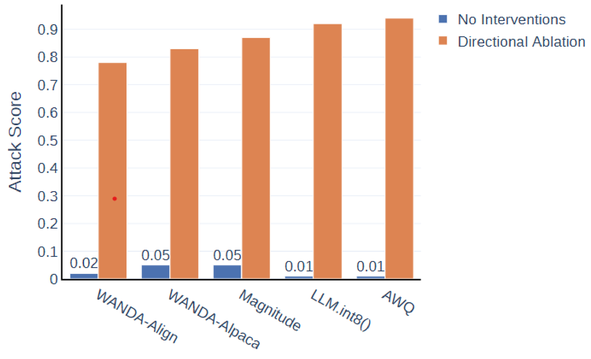
Vishnu Kabir Chhabra, Mohammad Mahdi Khalili



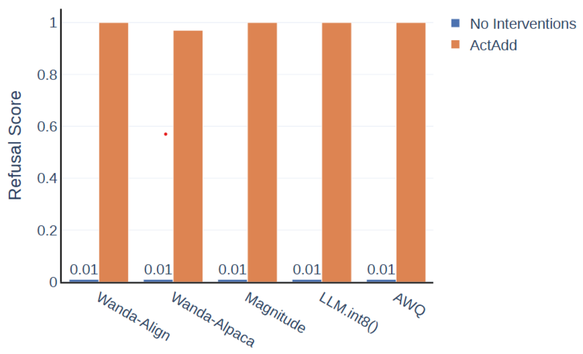
Pruning Changes Refusal Direction, Compression Doesn't

Type	Model	Method	l^c/l	i^c/i	Calibration Type
Pruning	Llama2-7b	Wanda	14/14	−5/ − 1	Alpaca
	Llama2-7b	Wanda	12/14	−5/ − 1	Align
	Llama2-7b	Magnitude	12/14	−5/ − 1	—
	Llama3-8b	Wanda	12/12	−5/ − 5	Alpaca
	Llama3-8b	Wanda	13/12	−5/ − 1	Align
Quantization	LLama2-7b	LLM.int8()	14/14	−1/ − 1	—
	LLama2-7b	AWQ	14/14	−1/ − 1	Pile
	LLAma3-8b	LLM.int8()	12/12	−5/ − 5	—
	LLAma3-8b	AWQ	12/12	−5/ − 5	Pile

Necessity Of Refusal Directions



Sufficiency Of Refusal Directions

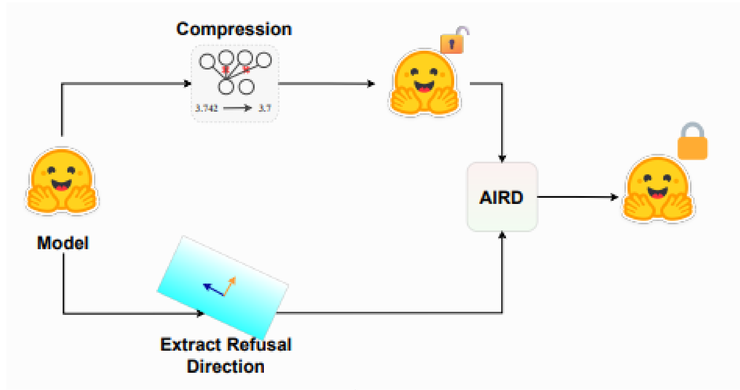


Pruning Breaks Safety!

Model	Method	$ASR^I_{\text{Adv-Decoding}}$	ASR_{Vanilla}	$ASR^\times_{\text{Adv-Decoding}}$	$ASR_{\text{Adv-Suffix}}$
Llama2	Base	0.006	0.16	0.27	0.09
Llama2	Wanda-Align	0.0	0.17	0.30	0.13
Llama2	Wanda-Alpaca	0.022	0.17	0.316	0.24
Llama2	Magnitude	0.01	0.6	0.496	0.35
Llama3	Base	0.054	0.01	0.046	0.01
Llama3	Wanda-Align	0.07	0.01	0.076	0.04
Llama3	Wanda-Alpaca	0.112	0.04	0.12	0.13

Quantization Doesn't Affect The Refusal Direction

Model	Method	Cosine Similarity
Llama2-7b	Wanda-Align	0.351
Llama2-7b	LLM.int8()	0.996
LLama2-7b	Wanda-Alpaca	0.539
Llama2-7b	AWQ	0.996
Llama2-7b	Magnitude	0.337
Llama3-8b	Wanda-Align	0.732
Llama3-8b	LLM.int8()	0.99
Llama3-8b	Wanda-Alpaca	0.902
LLama3-8b	AWQ	0.994



AIRD

Method: Consider a model M with a refusal direction $\mathbf{r}_i^{(l)}$ and the compressed model M^c with $\mathbf{r}_{i^c}^{(l^c)}$ as its refusal direction. We orthogonalize the weight matrices that project to the residual stream (attention output and MLP output) in layer l in the compressed model with respect to the refusal direction $\mathbf{r}_i^{(l)}$ and add it to the weight matrix as follows,

$$W^c_{l,new} \leftarrow W^c_l + \alpha \mathbf{r}_i^{(l)} (\mathbf{r}_i^{(l)})^\intercal W^c_l, \qquad (6)$$

Model	Method	Calibration	$ASR^I_{\text{Adv-Decoding}}$	ASR_{Vanilla}	$ASR^\times_{\text{Adv-Decoding}}$	$ASR_{\text{Adv-Suffix}}$
Llama2-7b	WANDA	Align	0%	41%(↓)	14%(↓)	15%(↓)
Llama2-7b	WANDA	Alpaca	10%(↓)	17%(↓)	12.5%(↓)	41%(↓)
Llama2-7b	Magnitude	—	40%(↓)	20%(↓)	2.4%(↑)	14.2%(↓)
Llama3-8b	WANDA	Align	22.3%(↓)	18.4%(↓)	0%	0%
Llama3-8b	WANDA	Alpaca	10.7%(↓)	33.3%(↓)	17.87%(↓)	16.6%(↓)

Model	RTE	ARC	BoolQ	Winogrande	HellaSwag
Llama2 Wanda-Align	68.0 / 68.5 (-0.5)	36.5 / 36.0 (+0.5)	76.5 / 76 (+0.5)	64.5 / 63.0 (+1.5)	54.0 / 54.0 (+0.0)
Llama2 Wanda-Alpaca	63.5 / 64.5 (-1.0)	41.5 / 40.5 (+1.0)	79.0 / 79.0 (+0.0)	66.0 / 66.5 (-0.5)	55.5/ 55.5 (+0.0)
Llama2 Magnitude	52.0 / 54.0 (-2.0)	34.5 / 34.0 (-0.5)	68.5 / 69.0 (-0.5)	61.5 / 63.0 (-1.5)	48.0/ 48.0 (+0.0)
Llama3 Wanda-Align	62.0 / 62.5 (-0.5)	43.5 / 44.5 (-0.5)	79.0 / 78.5 (+0.5)	70.0 / 71.0 (-1.0)	50.5 / 50.5 (+0.0)
Llama3 Wanda-Alpaca	62.5 / 62.5 (+0.0)	46.0 / 45.0 (+1.0)	82.0 / 82.0 (+0.0)	68.5 / 67.5 (+1.0)	51.5 / 51.5 (+0.0)

THE OHIO STATE UNIVERSITY
COLLEGE OF ENGINEERING