

# CoT Locality: Add/Lesion Effects (Layers 25-27)

Baseline Accuracy: 65.0%

$\Delta = \text{Intervention Accuracy} - \text{Baseline Accuracy}$

p-values: McNemar's Test | \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$  | Bold = Significant

Add ( $\alpha$ )



Lesion ( $\gamma$ )



Layer 25

Layer 26

Layer 27

$\Delta$  Answer Accuracy  
(Intervention – Baseline)

0.5

1.0

2.0

Parameter Value

0.5

1.0

2.0

Parameter Value

-0.2

-0.1

0.0

0.1

0.2