

## Layer 25: COT vs Answer Locality Comparison

Baseline Accuracy: 65.0% |  $\Delta$  = Intervention Accuracy – Baseline Accuracy  
P-values: McNemar's test | \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$  | Bold = Significant

