

Surgical Precision: Only Layer 10 Attention Succeeds

Component Type

Attention

MLP

Full Layer

Layer 8

Layer 9

Layer 10

Layer 11

Layer 12

Transformer Layer

2%

8%

100%
SUCCESS

5%

3%

0%

3%

15%

2%

0%

5%

10%

25%

8%

4%

Intervention Success Rate (%)

Intervention Results:	
✓	Layer 10 Attention: 100%
Δ	Layer 10 MLP: 25%
Δ	Layer 10 Full Layer: 15%
×	Other layers: <10%
Success threshold: >95%	
Total tested: 15 configs	