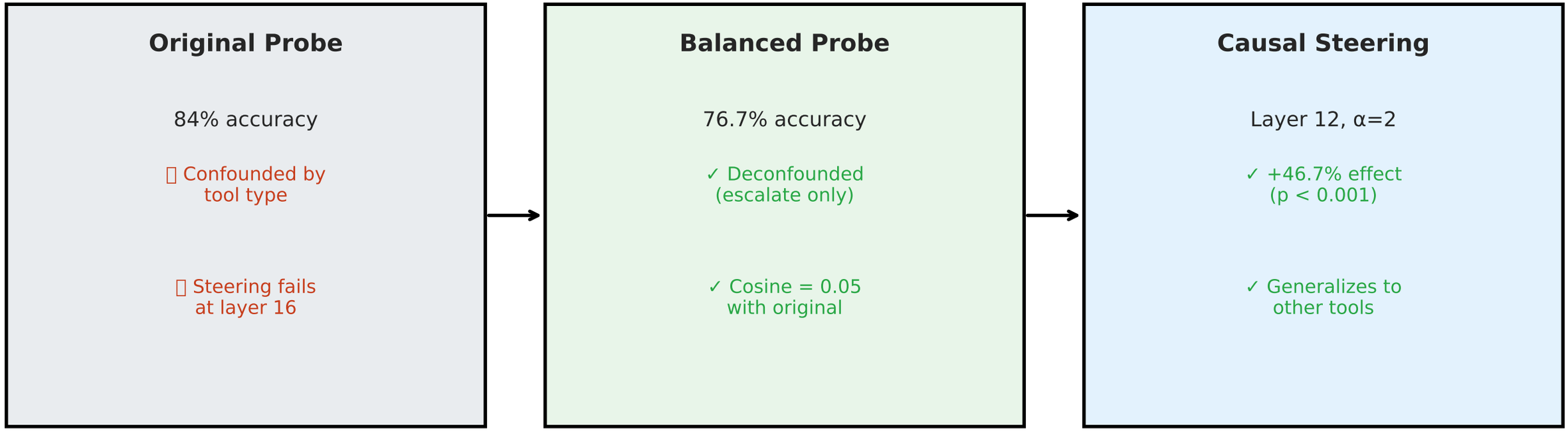
Causal Steering of Action-Grounding



Key Insight

Optimal probe layer (16) ≠ Optimal intervention layer (12)

Probes read post-decision states; interventions must target decision-making layers