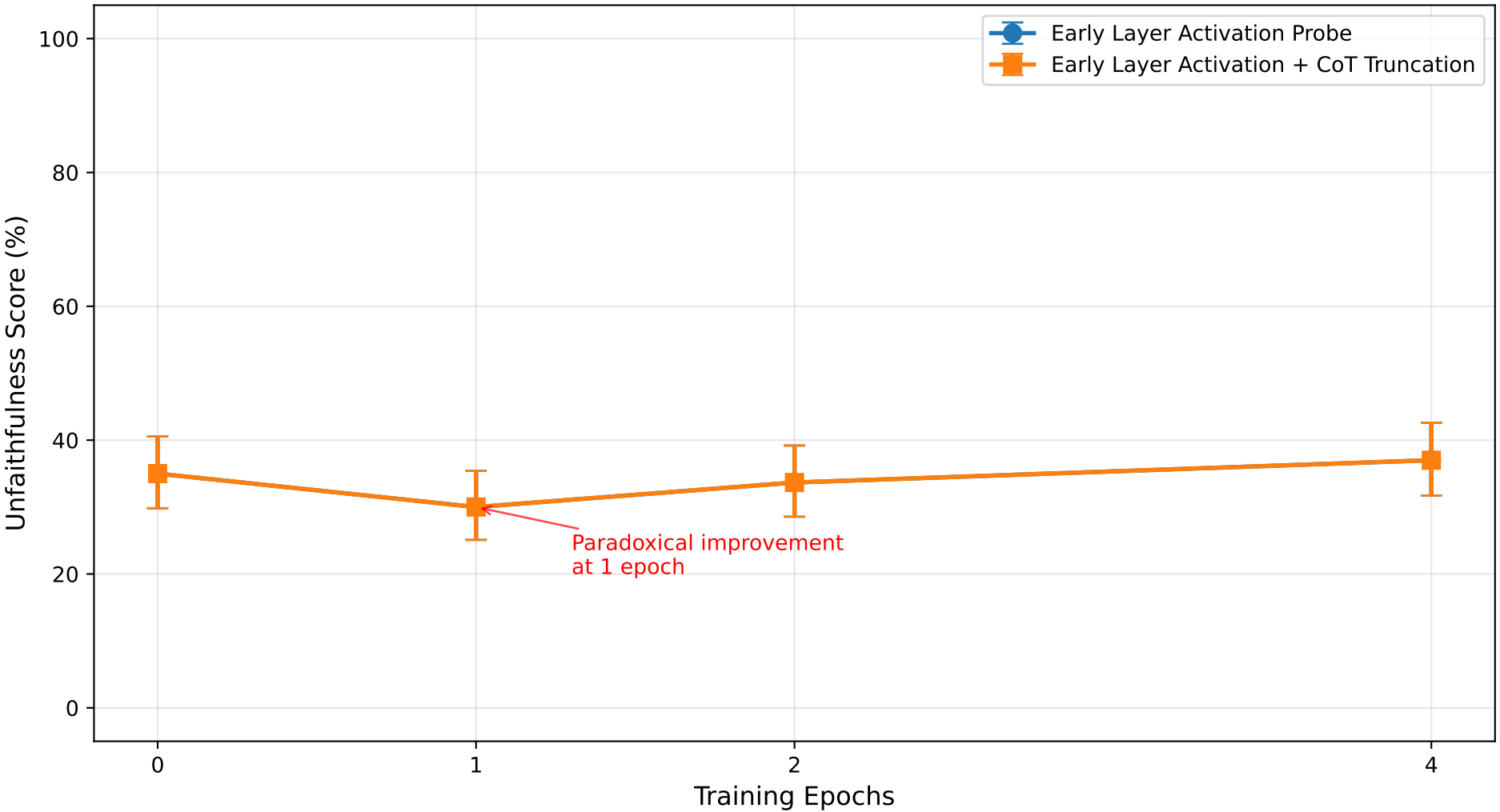
White-Box (Early Layer Activation Probing) vs Hybrid Detection Methods
Qwen3-4B, 20000 Documents
Early Layer Activation Probe
Early Layer Activation + CoT Truncation
Unfaithfulness Score (%)
100
80
60
40
20
0
Paradoxical improvement
at 1 epoch
0
1
2
4
Training Epochs