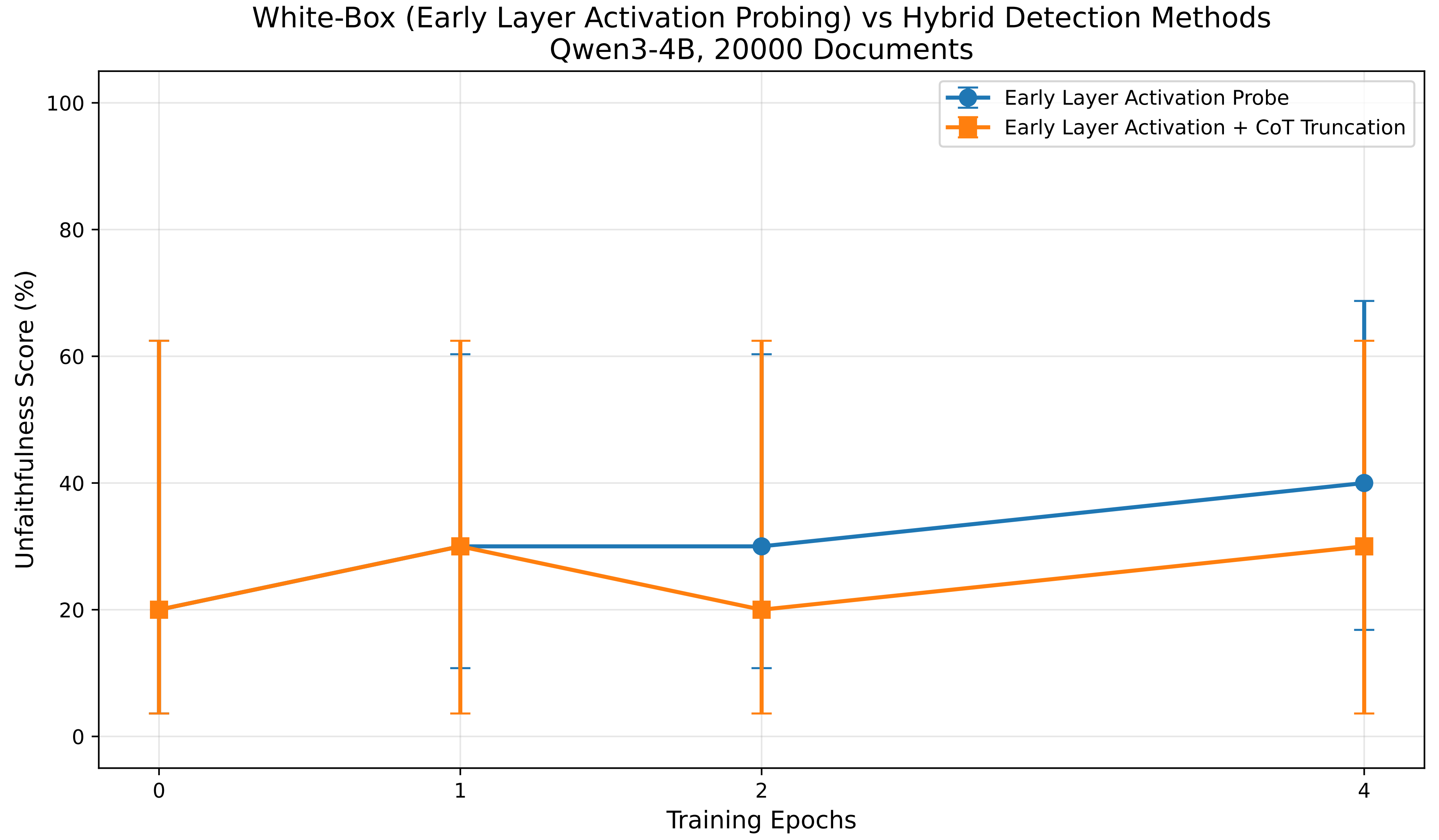
White-Box (Early Layer Activation Probing) vs Hybrid Detection Methods
Qwen3-4B, 20000 Documents
Early Layer Activation Probe
Early Layer Activation + CoT Truncation
Unfaithfulness Score (%)
0
20
40
60
80
100
Training Epochs
0
1
2
4