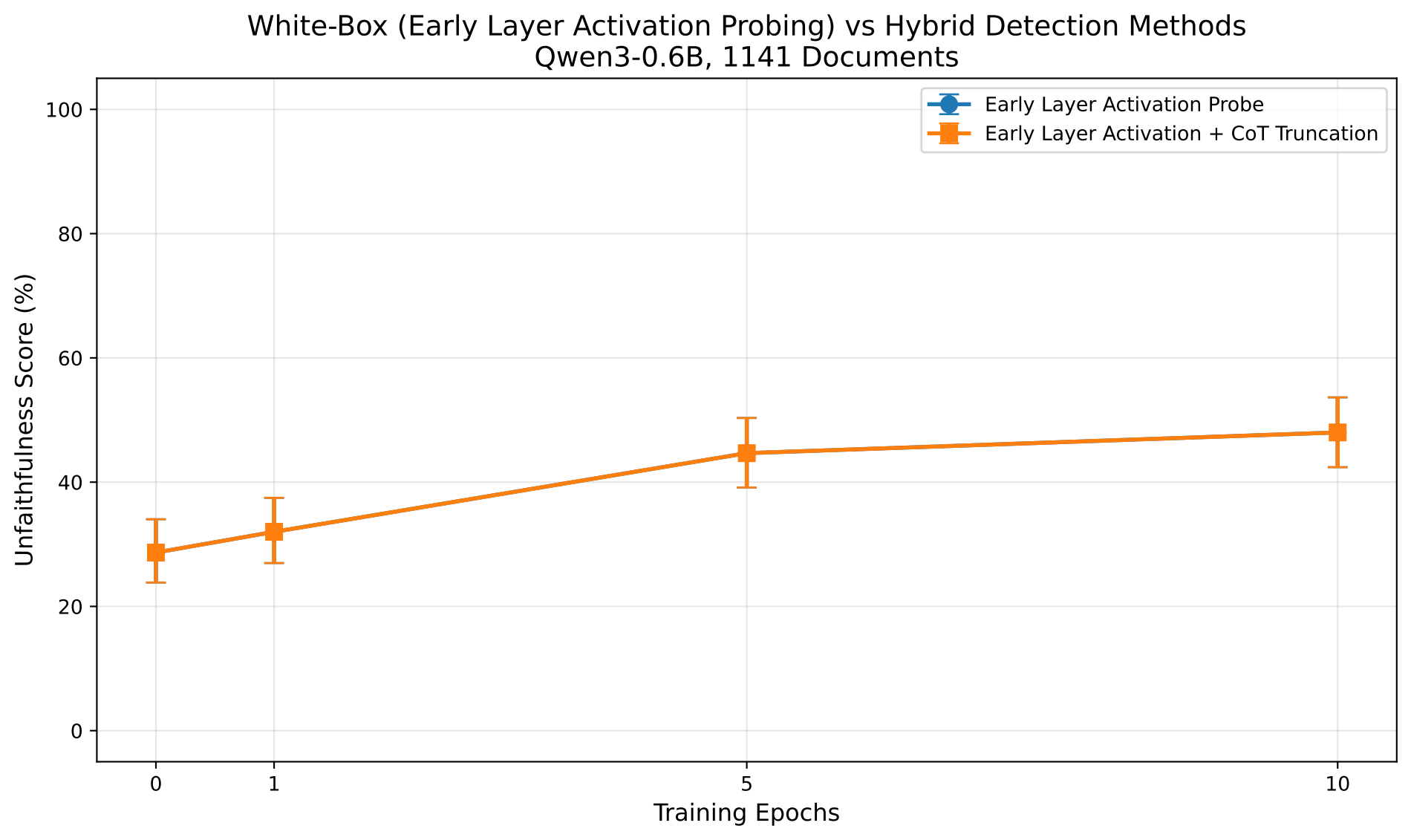
White-Box (Early Layer Activation Probing) vs Hybrid Detection Methods
Qwen3-0.6B, 1141 Documents
Early Layer Activation Probe
Early Layer Activation + CoT Truncation
Unfaithfulness Score (%)
100
80
60
40
20
0
0
1
5
10
Training Epochs