

White-Box (Early Layer Activation Probing) vs Hybrid Detection Methods  
Qwen3-4B, 20000 Documents

