

White-Box (Early Layer Activation Probing) vs Hybrid Detection Methods
Qwen3-0.6B, 20000 Documents

