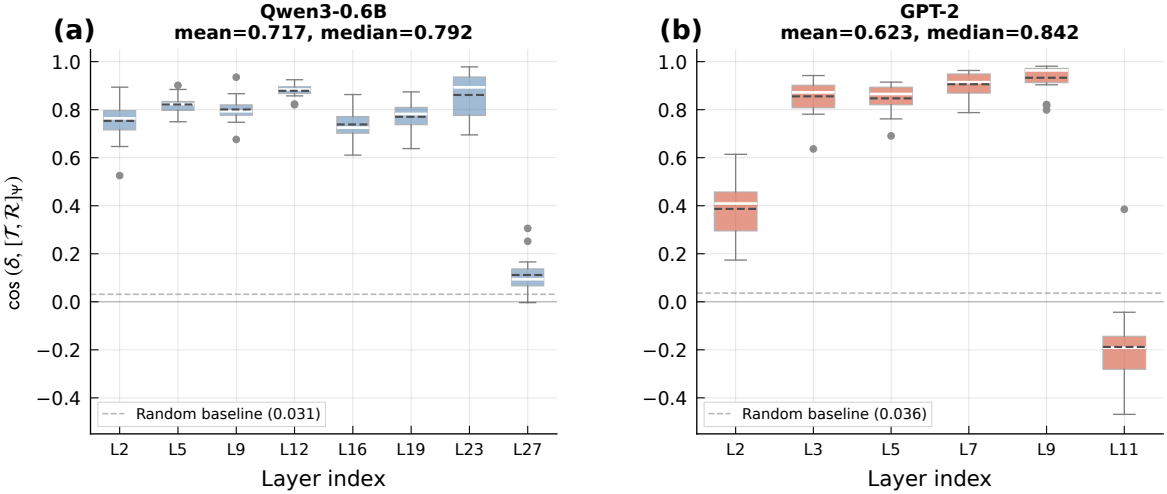
(a)
Qwen3-0.6B
mean=0.717, median=0.792
(b)
GPT-2
mean=0.623, median=0.842
$\cos(\delta, [\mathcal{T}, \mathcal{R}]_\Psi)$
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
Random baseline (0.031)
Random baseline (0.036)
L2
L5
L9
L12
L16
L19
L23
L27
L3
L7
L11
Layer index