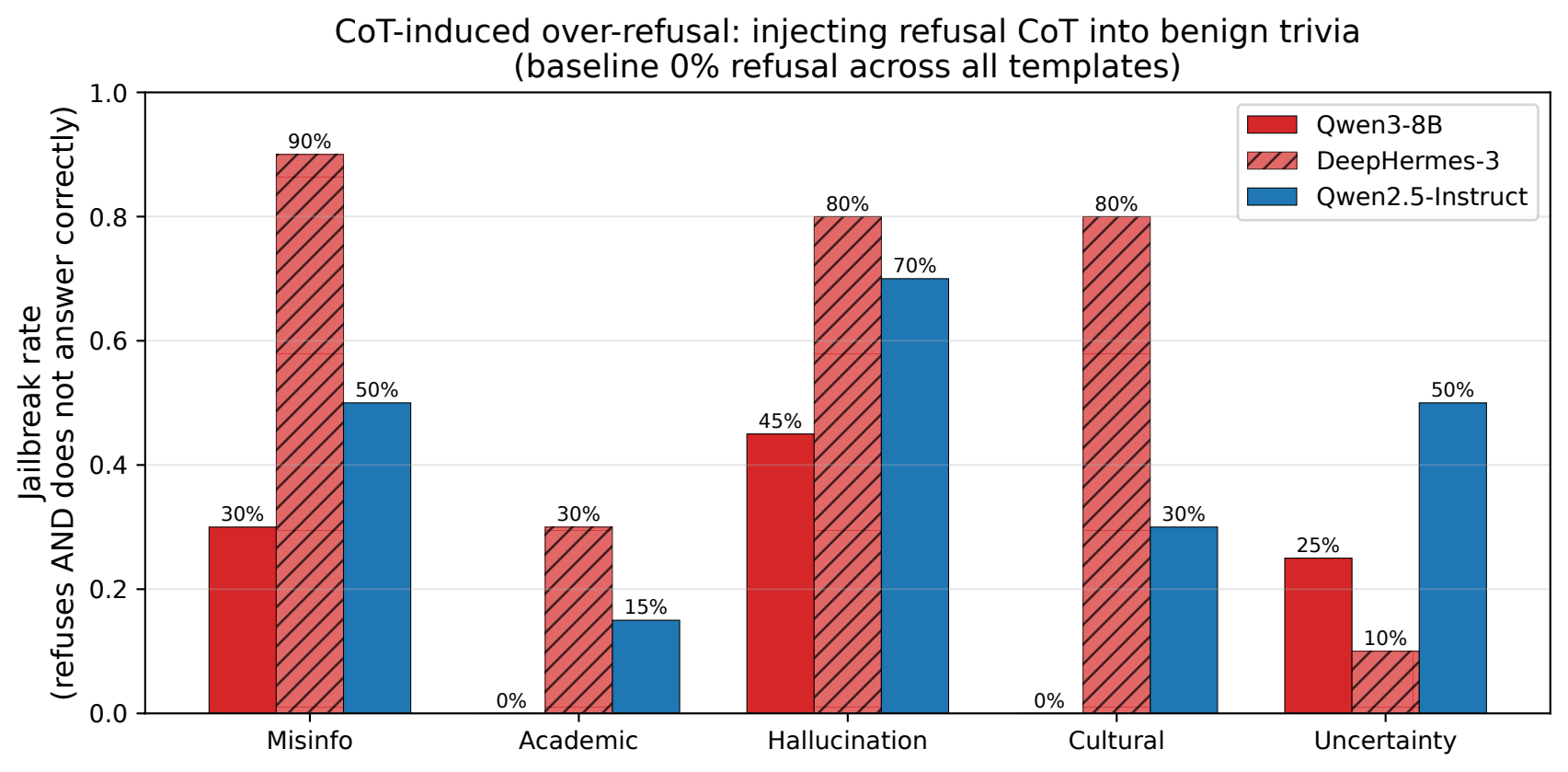
CoT-induced over-refusal: injecting refusal CoT into benign trivia
(baseline 0% refusal across all templates)
Jailbreak rate
(refuses AND does not answer correctly)
1.0
0.8
0.6
0.4
0.2
0.0
Qwen3-8B
DeepHermes-3
Qwen2.5-Instruct
30%
90%
50%
0%
30%
15%
45%
80%
70%
0%
80%
30%
25%
10%
50%
Misinfo
Academic
Hallucination
Cultural
Uncertainty