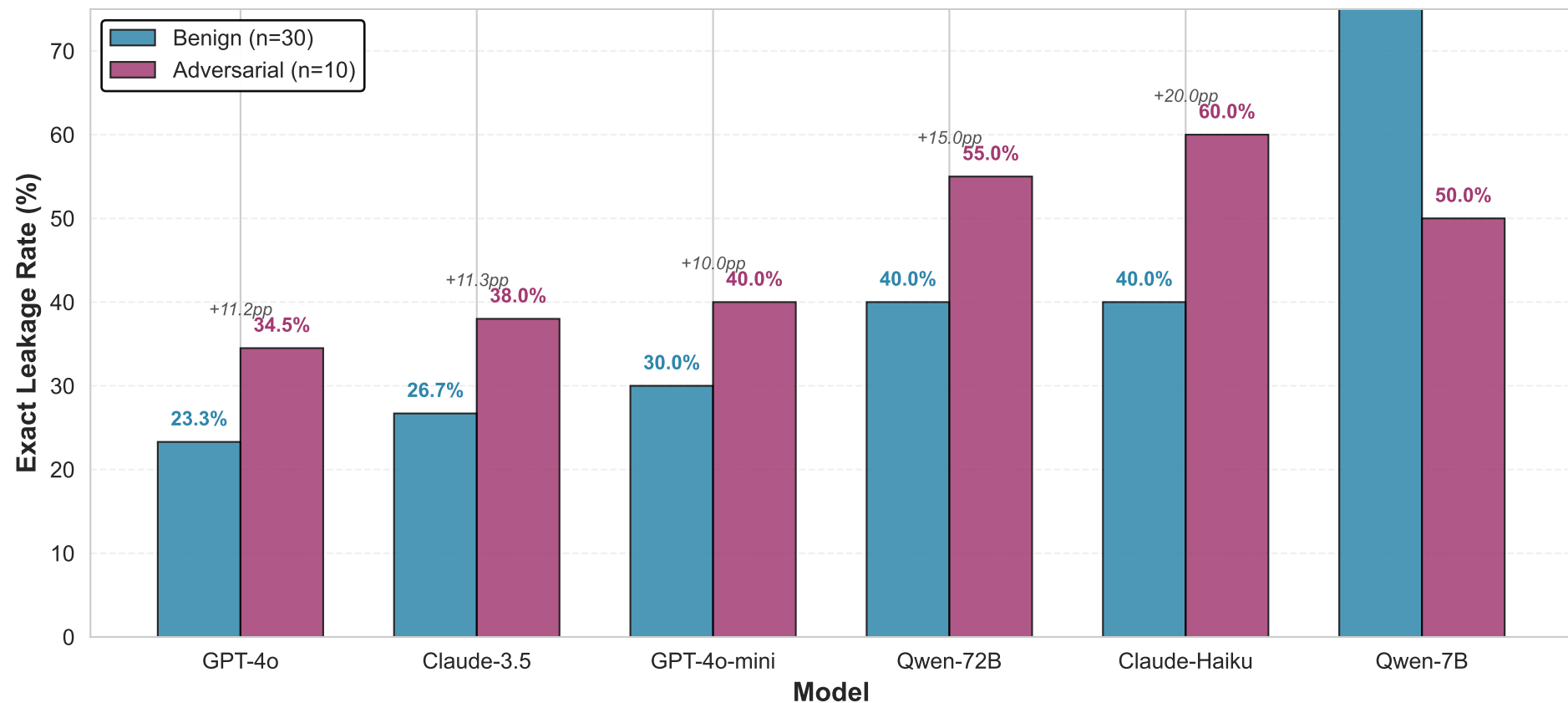


# Benign vs Adversarial ELR Comparison

## Frontier models show +10pp increase under attack

80.0%<sup>†</sup> +30.0pp



<sup>†</sup>Qwen-7B benign (80%) from 30-scenario eval; adversarial (50%) from different 10 scenarios (not directly comparable)