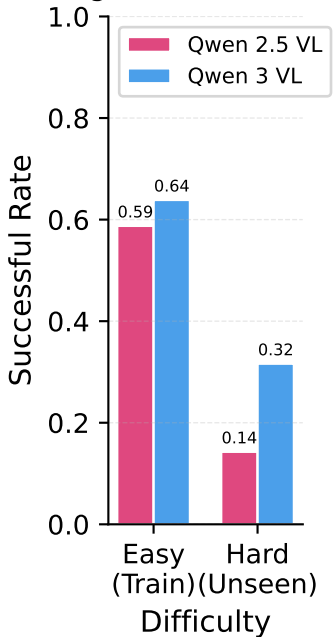


SFT Generalization to Harder Variants

Avg. Successful Rate



Per-Task Successful Rate

