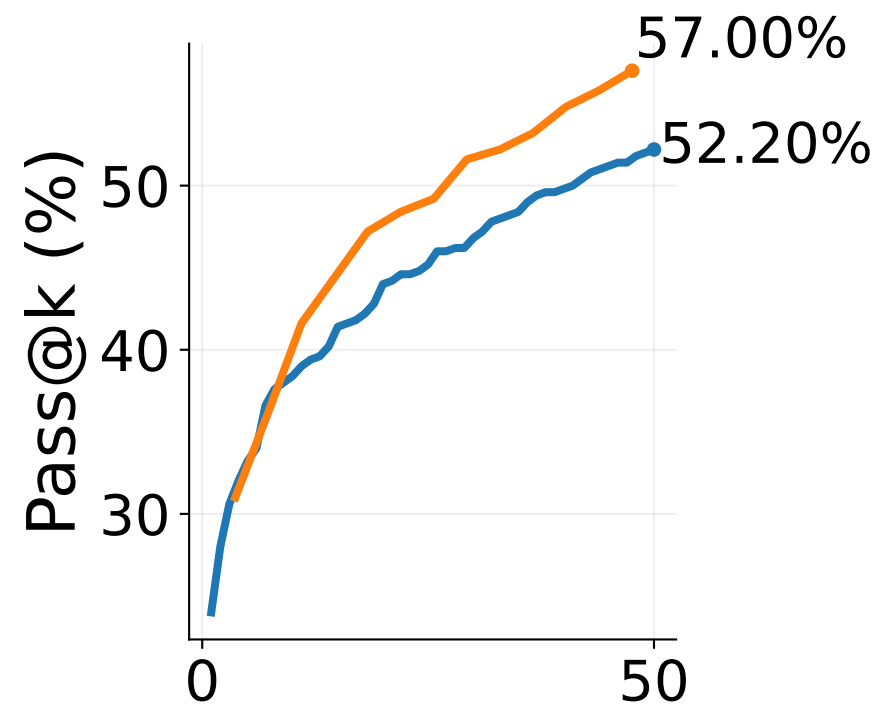
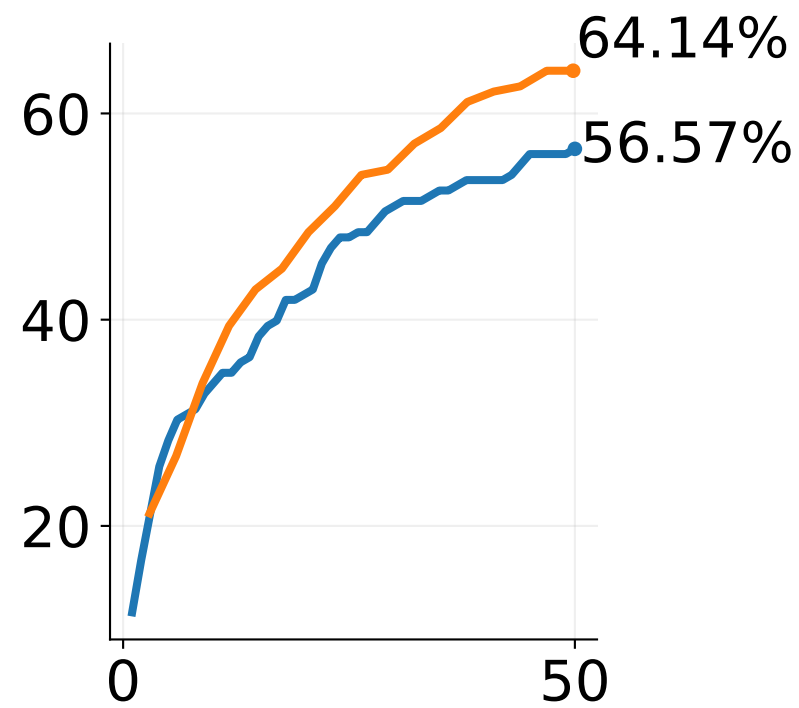


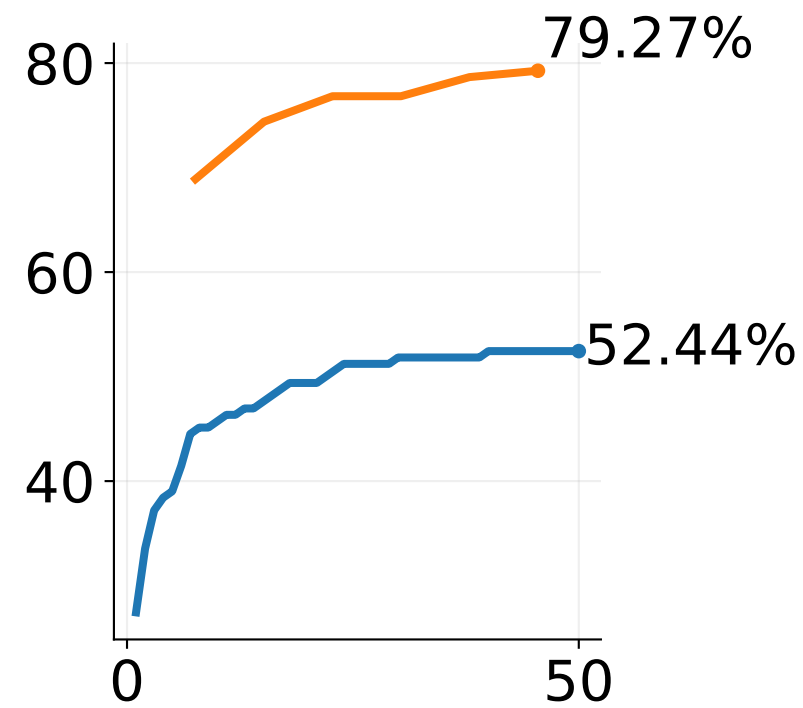
MATH



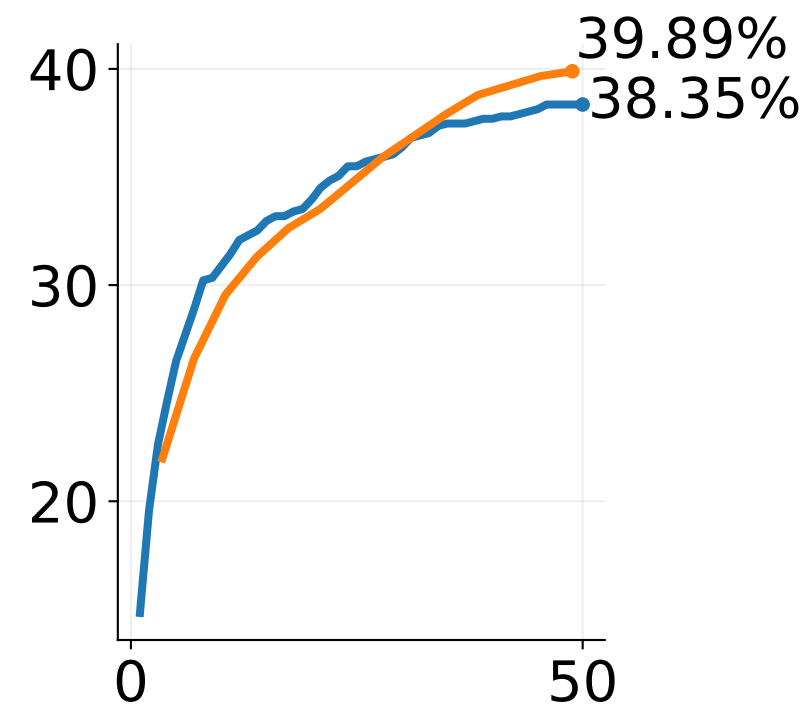
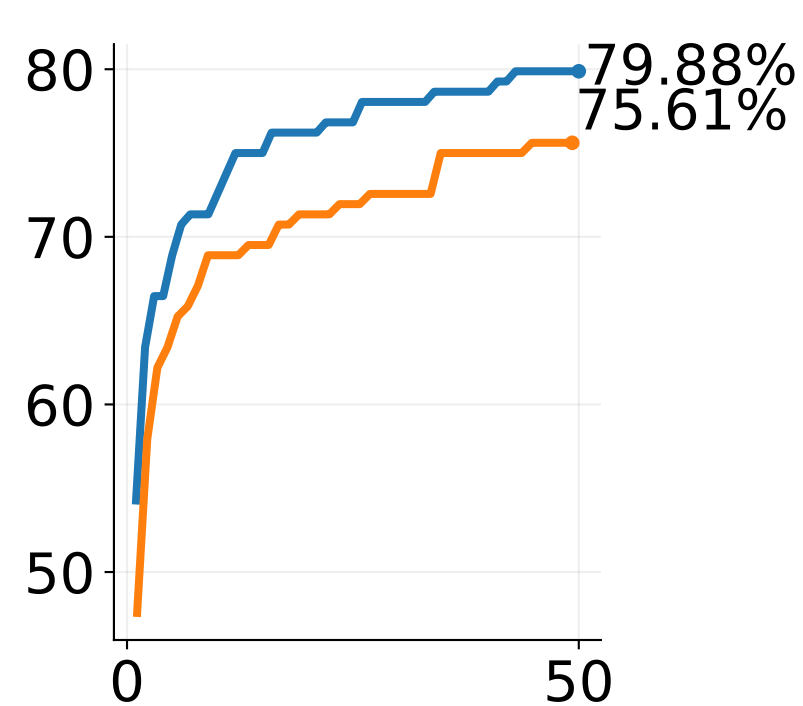
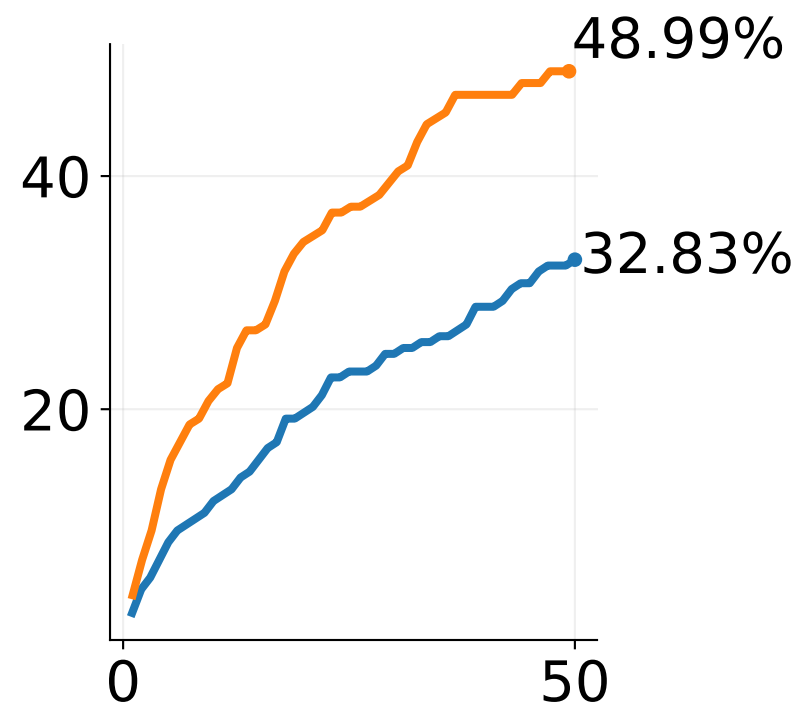
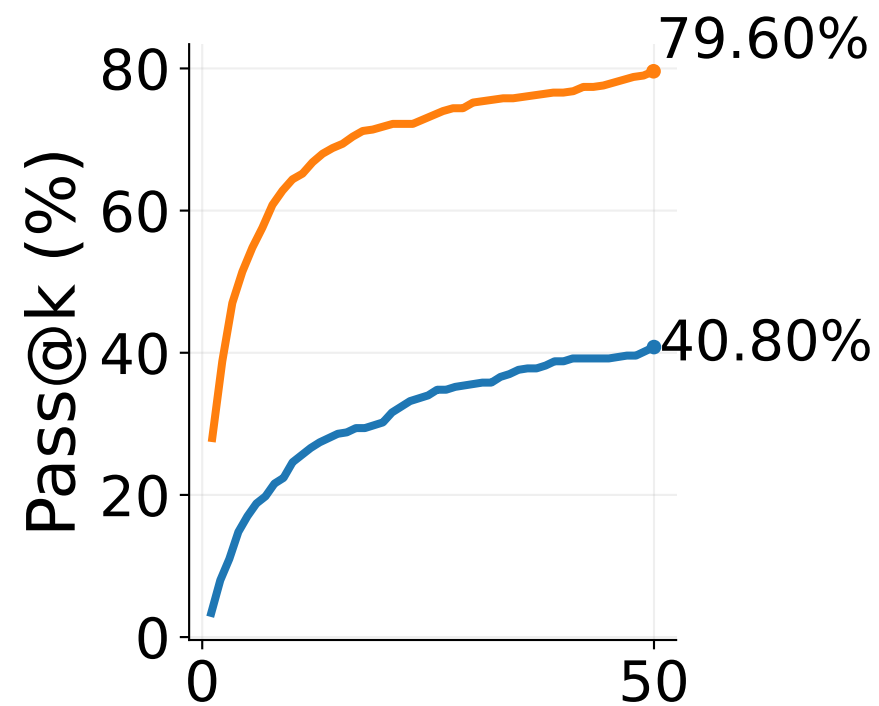
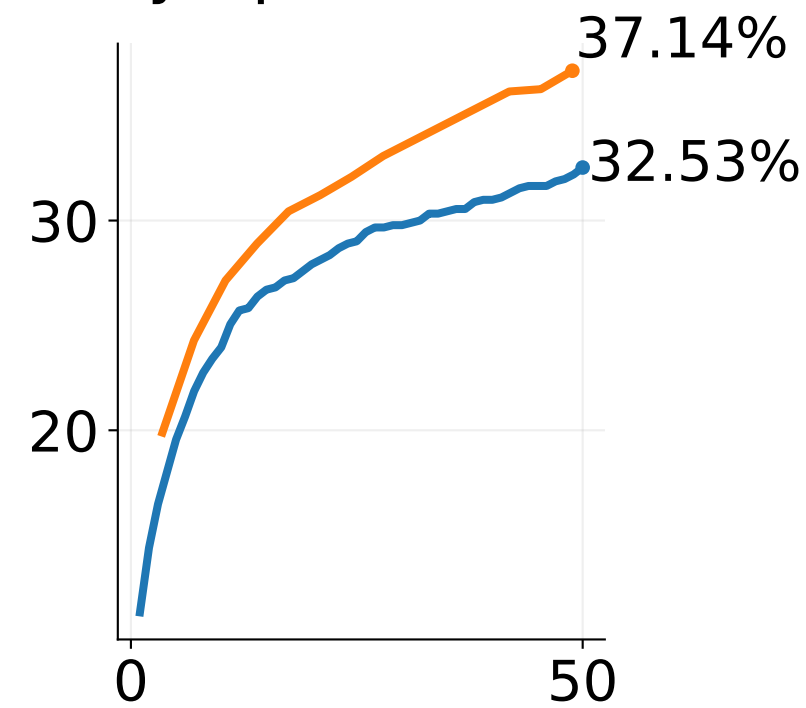
GPQA-Diamond



HumanEval



Olympiad Bench



k (Number of Attempts)

k (Number of Attempts)

k (Number of Attempts)

k (Number of Attempts)

— Traditional Repeated Sampling — GUIDEDSAMPLING