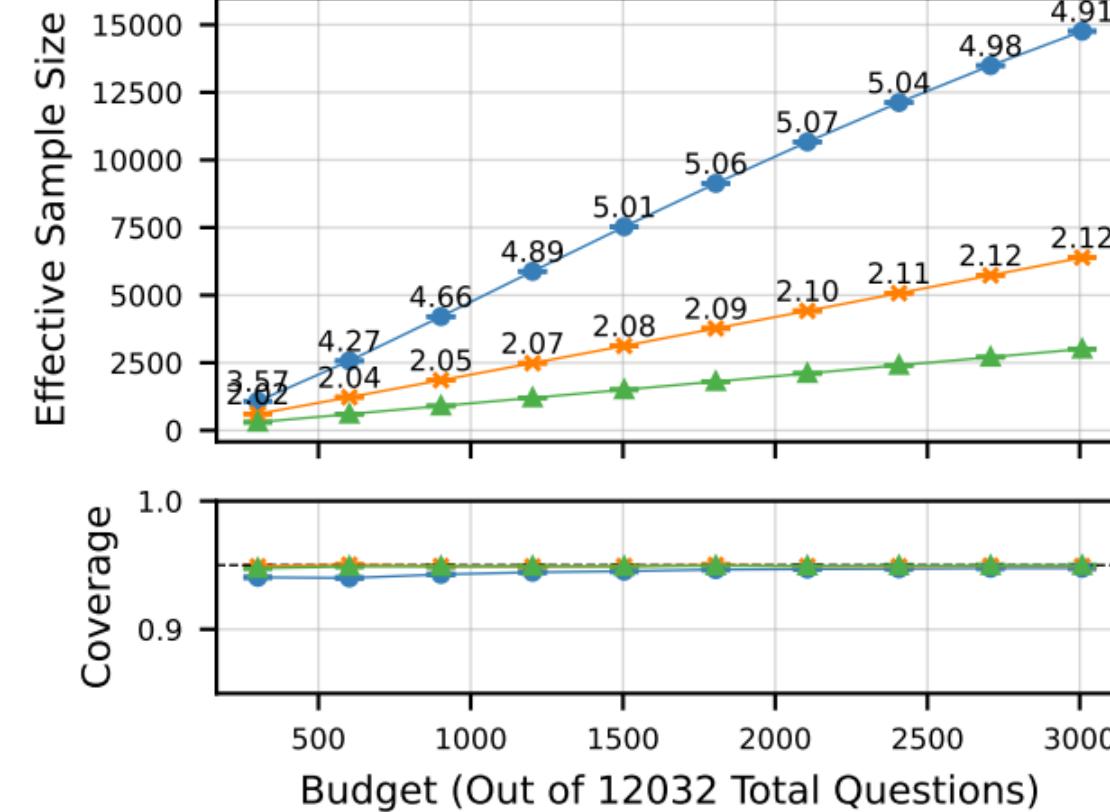
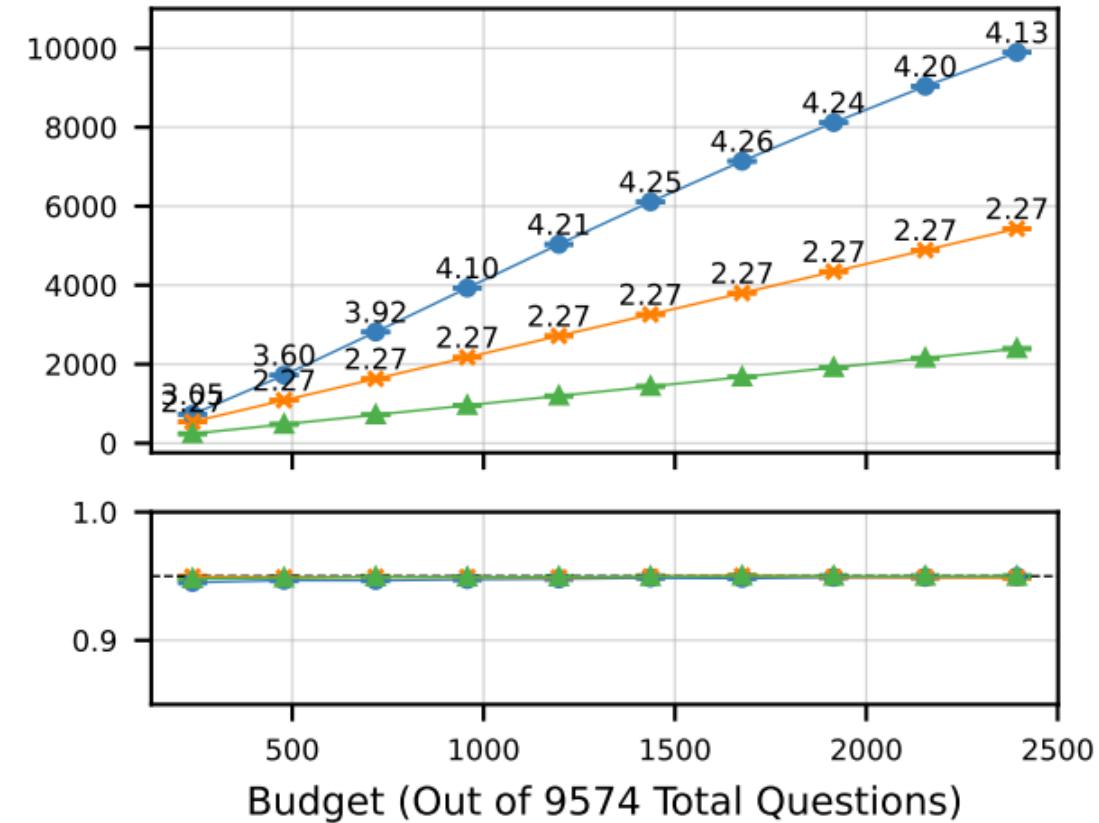


MMLU-Pro



BBH+GPQA+IFEval+MATH+MuSR



—●— FAQ   —★— Best Baseline (Post-hoc Per Budget)   —▲— Uniform   ----- 95% Coverage (Bottom Subplots)