



Figure 1: Weaker surrogates can effectively support effective data acquisition for active testing. We compare uniform-random testing against active testing using the few-shot Phi-2 model as the surrogate (LURE-Phi-2) and the cross-entropy acquisition function.