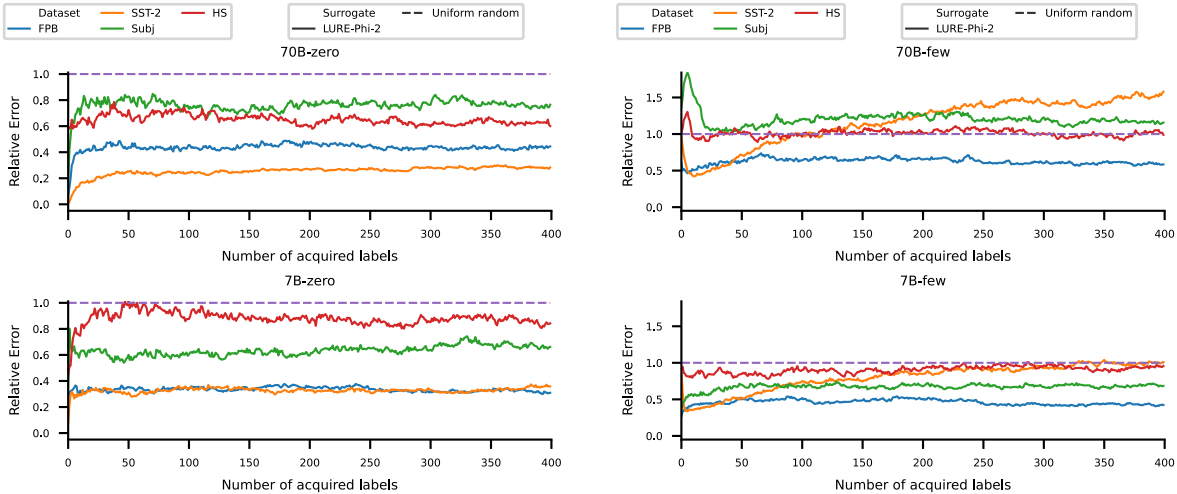

Figure 1: Relative error for zero-shot (a) and few-shot (b) evaluation, i.e. ratio of the target's risk estimation against uniform-random risk-estimation error, with Phi-2 as the surrogate and using the cross-entropy acquisition function. LURE-Phi-2 is consistently more effective than uniform random for 7B-few evaluation. For 70B-few evaluation, results are comparable to those of uniform sampling.