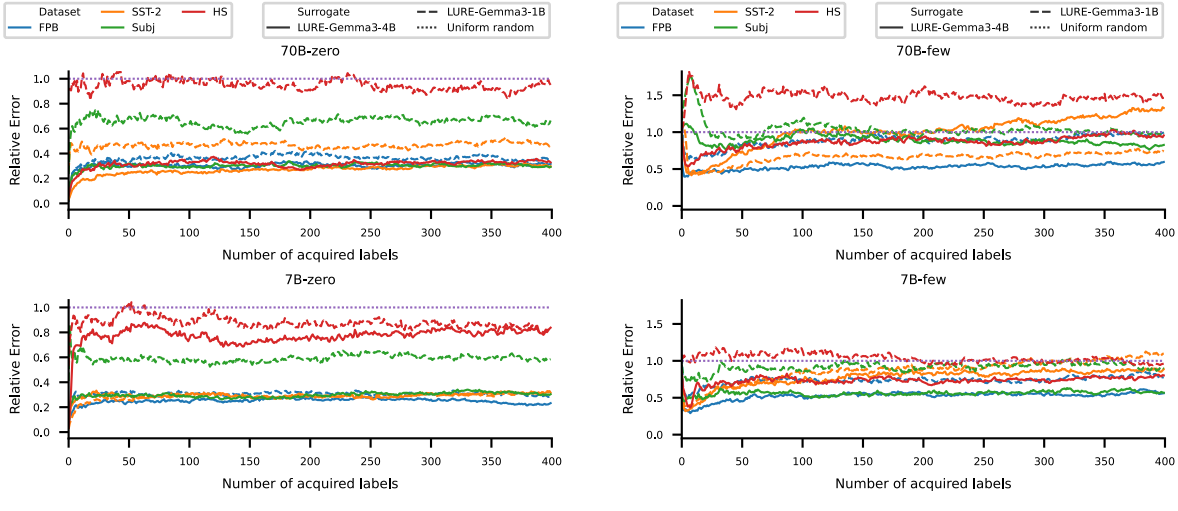

Figure 2: Relative error for zero-shot (a) and few-shot (b) evaluation, i.e. ratio of the target's risk estimation against uniform-random risk-estimation error, with Gemma-3 as the surrogate and using the cross-entropy acquisition function. Despite differences in architectures and Gemma-3's small size, active testing outperforms uniform sampling.