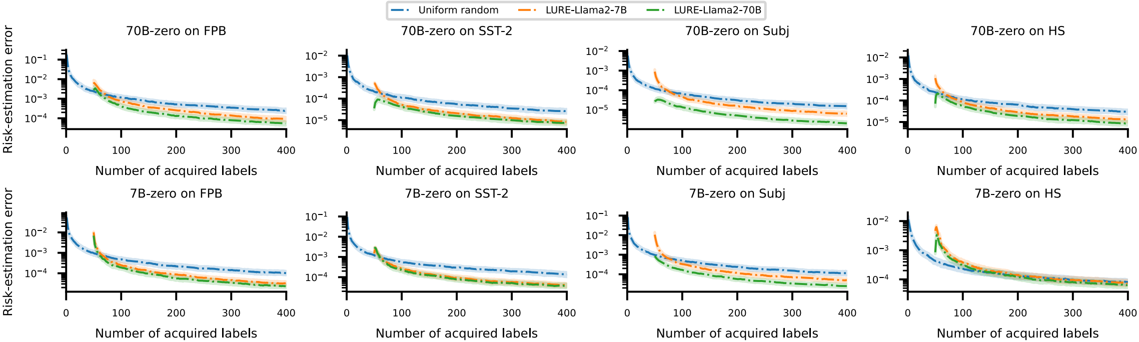

Figure 5: We add the context budget for the few-shot surrogate into the overall labeling budget for zero-shot evaluation. We show that active testing still achieves more precise evaluation than uniform sampling with a fixed labeling budget.