Figure 6: Weaker surrogates can effectively support effective data acquisition for active testing. We compare uniform-random testing against active testing using the few-shot Gemma-3 1B and 4B models as the surrogate (LURE-Gemma3-1B and Gemma3-4B, respectively) and the cross-entropy acquisition function.