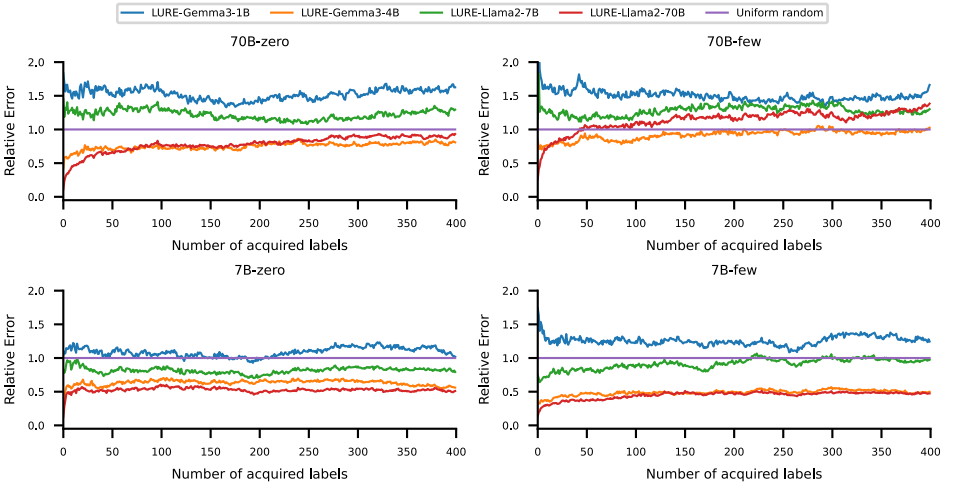

Figure 3: Relative error of active evaluation on the MMLU benchmark. Few-shot models use the same 5 in-context examples. For Llama-2 7B evaluation, active testing achieves higher precision than uniform sampling in most cases. Evaluating Llama-2 70B proves more challenging, where active testing gracefully degrades.