Closed-source models/LLM-as-a-judge experiment

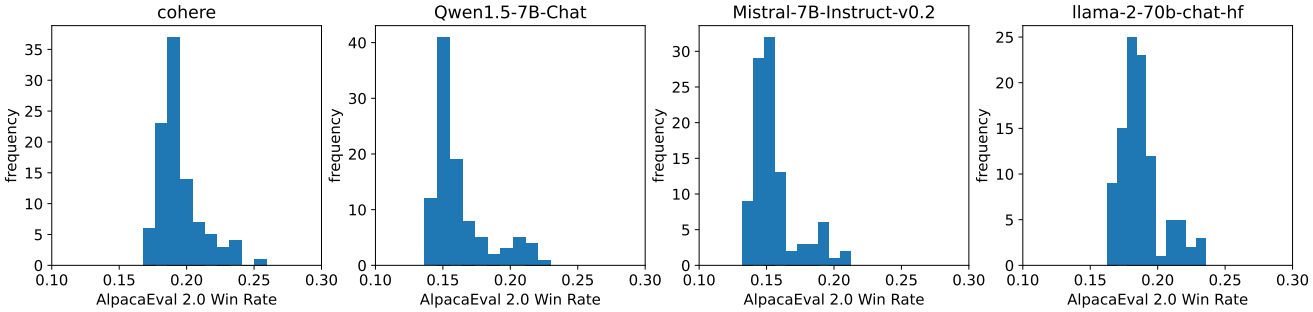

Figure 1: Verifying the sensitivity of the "LLM-as-a-judge" framework for LLM evaluation. We generated 100 prompt variations with ChatGPT and presented them to the closed-source judge (GPT-4o-mini) to assess performance sensitivity in the AlpacaEval 2.0 win rate. We can see that the judge is sensitive to the prompt template when rating models, which can cause issues such as varying rankings and poor estimates of performance.

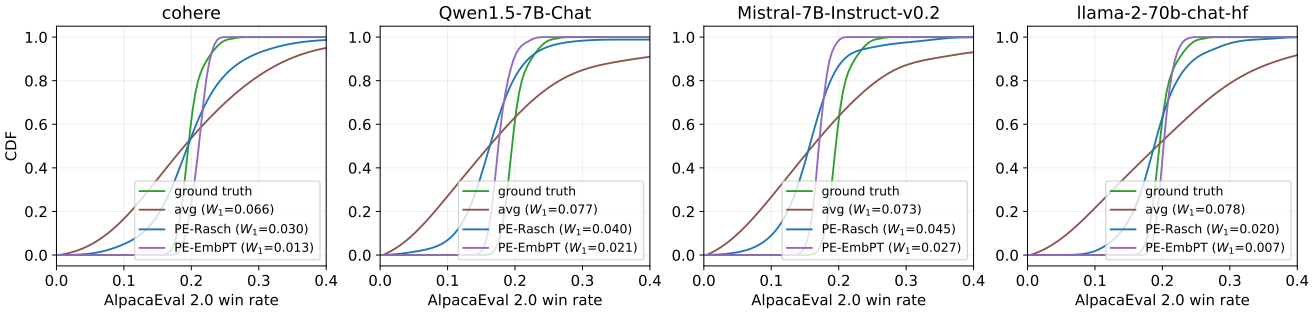

Figure 2: Estimating the performance variability in an LLM-as-a-judge benchmark (AlpacaEval 2.0). We generated 100 prompt variations with ChatGPT and presented them to the closed-source judge (GPT-4o-mini) to assess performance sensitivity. Evaluating four LLMs (cohere, Qwen1.5-7B-Chat, Mistral-7B-Instruct-v0.2, llama-2-70b-chat-hf) with $\approx 2\%$ of the total evaluations (1.6k/80.5k), we found that PromptEval offers superior performance in estimation error (Wasserstein distance W_1 between estimate and ground truth) compared to the baseline.

Experiments with less prompt variations

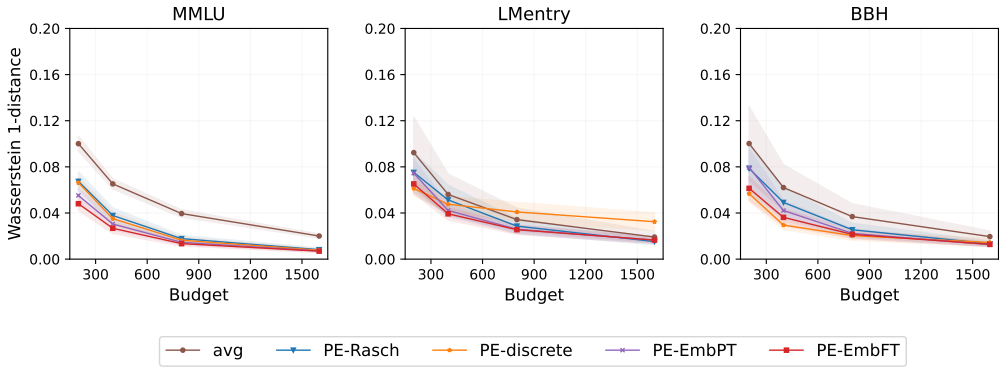

Figure 3: Performance estimation quality when we keep only 20% of the prompts variations for each task. In summary, PromptEval still does beat the baseline. However, the gap between PromptEval and the baseline has shortened due to fewer variations.