The qualitative evaluation results folder contains the detailed statistical comparison for every model pair per dataset in addition to Table 3. For each dataset, we computed the statistical results every possible model pairs using the two-sided Wilcoxon rank-sum test mentioned in the Evaluation section of the main manuscript and S2.1 in the supplementary material.

The results per dataset are summarized in each spreadsheet. Each row in the spreadsheet contains the statistical results of a model pair. The model name is structured as [dataset]\_[model]\_[setting]. The setting consists of zs for zero-shot, os for one-shot, 5s for five-shot, and finetuned for fine-tuning. For instance, BC5CDR\_Chemical\_gpt3.5\_zs is GPT-3.5 on the BC5CDR Chemical dataset under zero-shot. The statistical results consists of means, standard deviations, confidence internals, and P-values.