Stability of Dataset Comparisons for wig+100 (p=0.75) **Dataset Pair** beir-webis-touche2020 vs msmarco-passage beir-webis-touche2020 vs msmarco-subsample 0.25 msmarco-passage vs msmarco-subsample 0.20 Swap Probability 0.15 0.10 0.05 0.00

Average Performance Delta