Stability of Dataset Comparisons for wig+100 (p=0.25) 0.40 **Dataset Pair** beir-webis-touche2020 vs msmarco-passage beir-webis-touche2020 vs msmarco-subsample 0.35 msmarco-passage vs msmarco-subsample 0.30 Swap Probability 0.10 0.05 0.00 3

Average Performance Delta