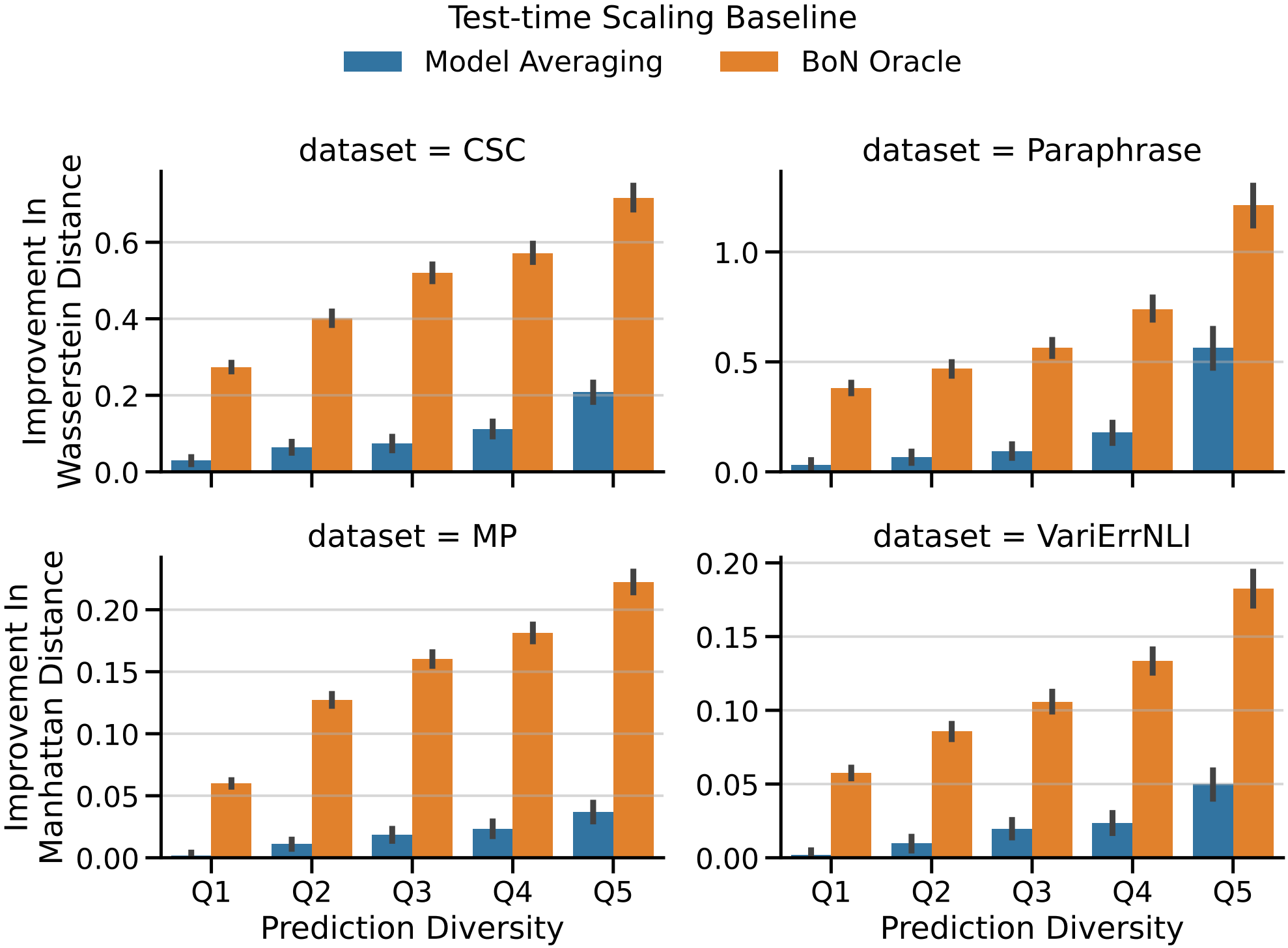
Test-time Scaling Baseline
Model Averaging
BoN Oracle
dataset = CSC
dataset = Paraphrase
dataset = MP
dataset = VariErrNLI
Improvement In Wasserstein Distance
Improvement In Manhattan Distance
0.0
0.2
0.4
0.6
0.5
1.0
0.00
0.05
0.10
0.15
0.20
Q1
Q2
Q3
Q4
Q5
Prediction Diversity