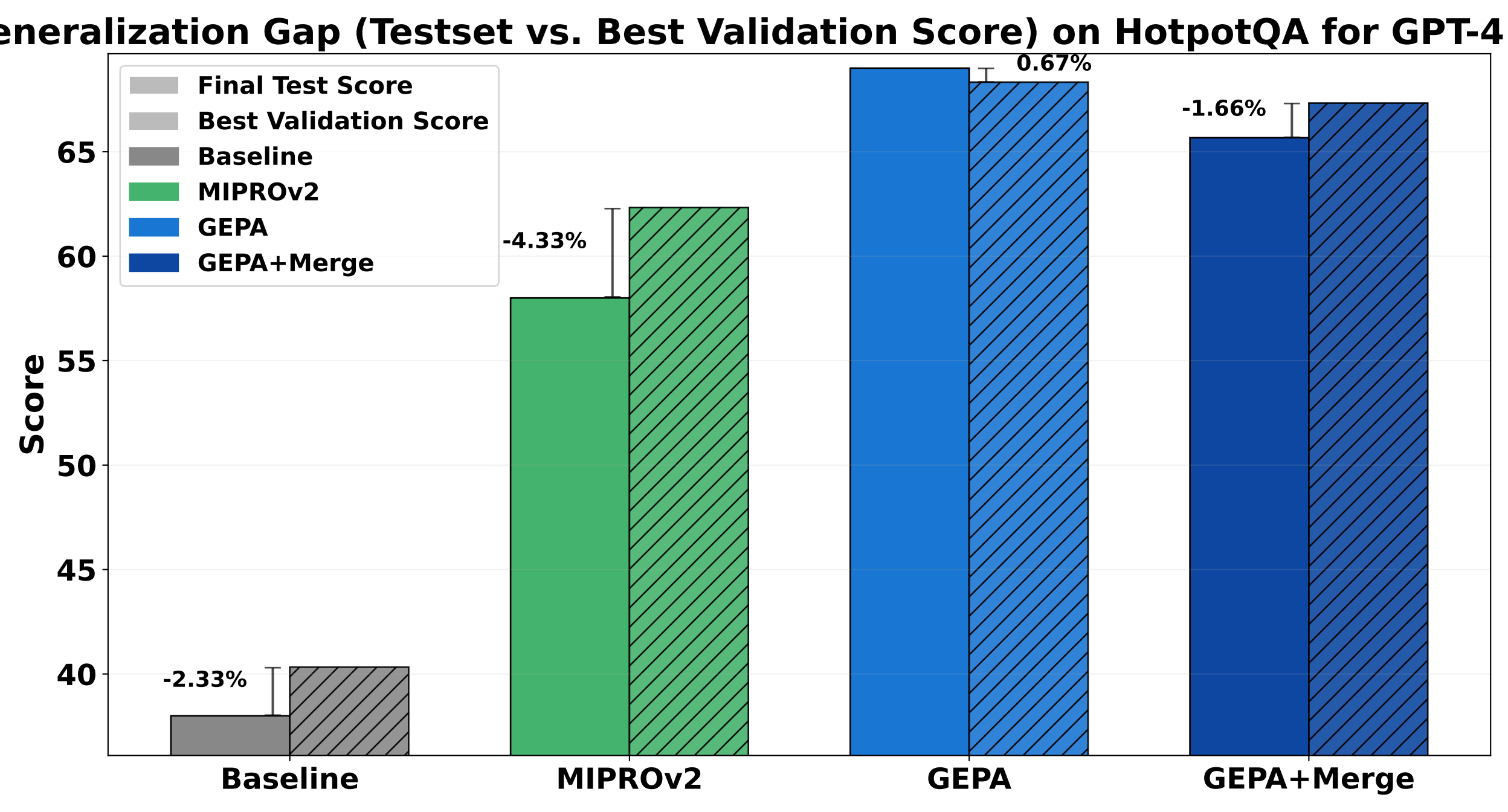
eneralization Gap (Testset vs. Best Validation Score) on HotpotQA for GPT-4
Final Test Score
Best Validation Score
Baseline
MIPROv2
GEPA
GEPA+Merge
Score
65
60
55
50
45
40
-2.33%
-4.33%
0.67%
-1.66%
Baseline
MIPROv2
GEPA
GEPA+Merge