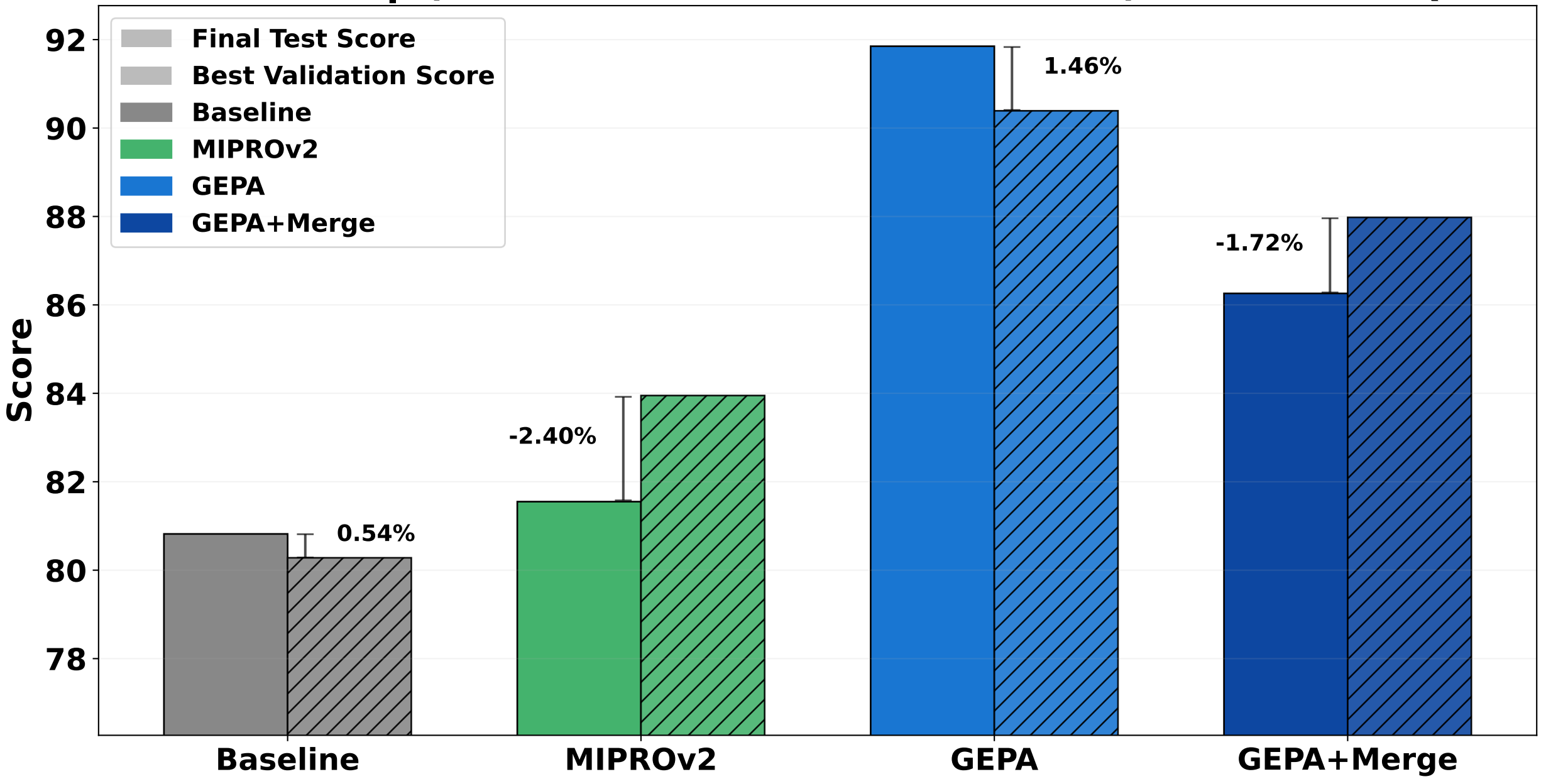
Generalization Gap (Testset vs. Best Validation Score) on PUPA for Qwen3
Final Test Score
Best Validation Score
Baseline
MIPROv2
GEPA
GEPA+Merge
Score
92
90
88
86
84
82
80
78
0.54%
-2.40%
1.46%
-1.72%
Baseline
MIPROv2
GEPA
GEPA+Merge