Generalization Gap (Testset vs. Best Validation Score) on HotpotQA for Qwen

Legend:
- Final Test Score
- Best Validation Score
- Baseline
- MIPROv2
- GEPA
- GEPA+Merge

Baseline: -4.00%
MIPROv2: -2.34%
GEPA: 2.33%
GEPA+Merge: 2.33%