Generalization Gap (Testset vs. Best Validation Score) on AIME2025 for GPT-4

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

Baseline: 2.66%
MIPROv2: -2.00%
GEPA: -0.67%