Qwen3-4B-Instruct 2507: Baseline vs Iterative Refinement

Legend:
- Baseline
- Iter. Refinement (r=3, old prompt)
- Iter. Refinement (r=12, old prompt)
- Iter. Refinement (r=12, new prompt)

HMMT Feb 2025: Baseline 28.75, Iter. Refinement (r=3, old prompt) 31.4, Iter. Refinement (r=12, old prompt) 32.5, Iter. Refinement (r=12, new prompt) 45.4

HMMT Nov 2025: Baseline 41.0, Iter. Refinement (r=3, old prompt) 44.3, Iter. Refinement (r=12, old prompt) 45.3, Iter. Refinement (r=12, new prompt) 56.5

IMO-Answerbench: Baseline 35.56, Iter. Refinement (r=3, old prompt) 38.6, Iter. Refinement (r=12, old prompt) 41.2, Iter. Refinement (r=12, new prompt) 47.2