Qwen3-8B: Baseline vs Reasoning Cache vs Iterative Refinement

Legend:
- Baseline (t=16384)
- Reasoning Cache (max rounds=12)
- Iter. Refinement (max rounds=12, new prompt)

Accuracy (%)

HMMT Feb 2025: 29.7, 43.4, 46.3
HMMT Nov 2025: 35.9, 50.3, 51.6
IMO-Answerbench: 30.3, 47.3, 44.9