GridPuzzle-GPT
gpt-4o-mini

Macro-F1 vs Reasoning-chain length (bin center)

Entail-Prev
Entail-Raw
Stability-0.95