GridPuzzle-Llama
gpt-4o-mini

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

Legend:
- Entail-Prev
- Entail-Raw
- Stability-0.95