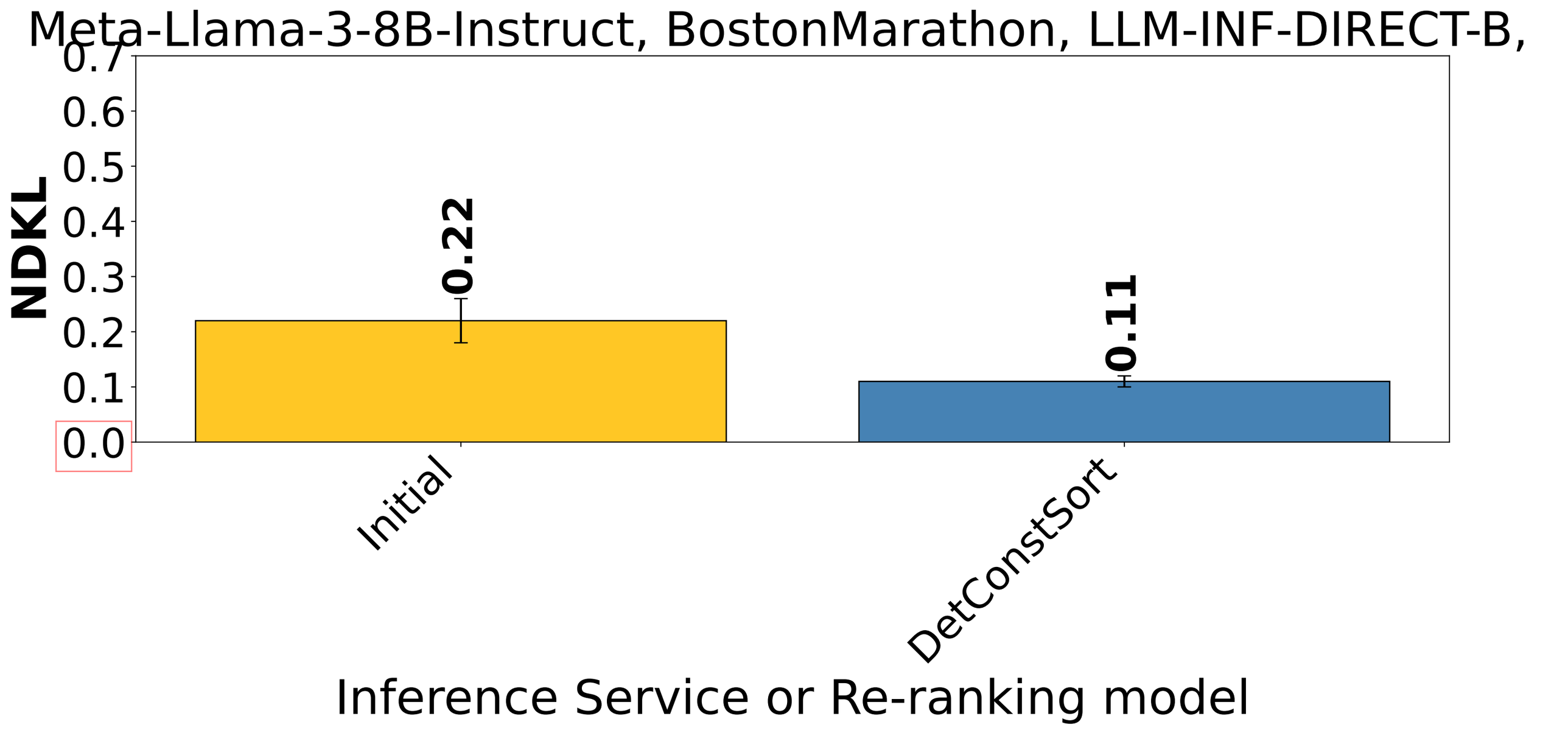
Meta-Llama-3-8B-Instruct, BostonMarathon, LLM-INF-DIRECT-B,
NDKL
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.22
0.11
Initial
DetConstSort
Inference Service or Re-ranking model