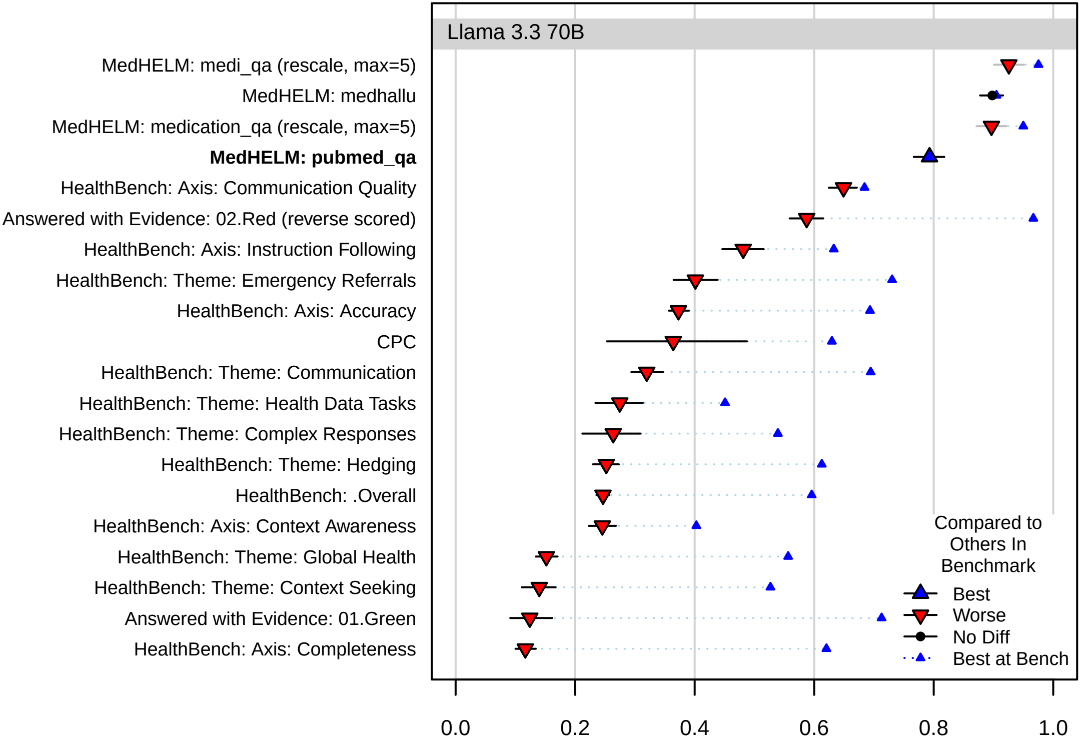
Llama 3.3 70B
MedHELM: medi_qa (rescale, max=5)
MedHELM: medhallu
MedHELM: medication_qa (rescale, max=5)
MedHELM: pubmed_qa
HealthBench: Axis: Communication Quality
Answered with Evidence: 02.Red (reverse scored)
HealthBench: Axis: Instruction Following
HealthBench: Theme: Emergency Referrals
HealthBench: Axis: Accuracy
CPC
HealthBench: Theme: Communication
HealthBench: Theme: Health Data Tasks
HealthBench: Theme: Complex Responses
HealthBench: Theme: Hedging
HealthBench: .Overall
HealthBench: Axis: Context Awareness
HealthBench: Theme: Global Health
HealthBench: Theme: Context Seeking
Answered with Evidence: 01.Green
HealthBench: Axis: Completeness
Compared to Others In Benchmark
▲ Best
▼ Worse
● No Diff
▲ Best at Bench
0.0
0.2
0.4
0.6
0.8
1.0