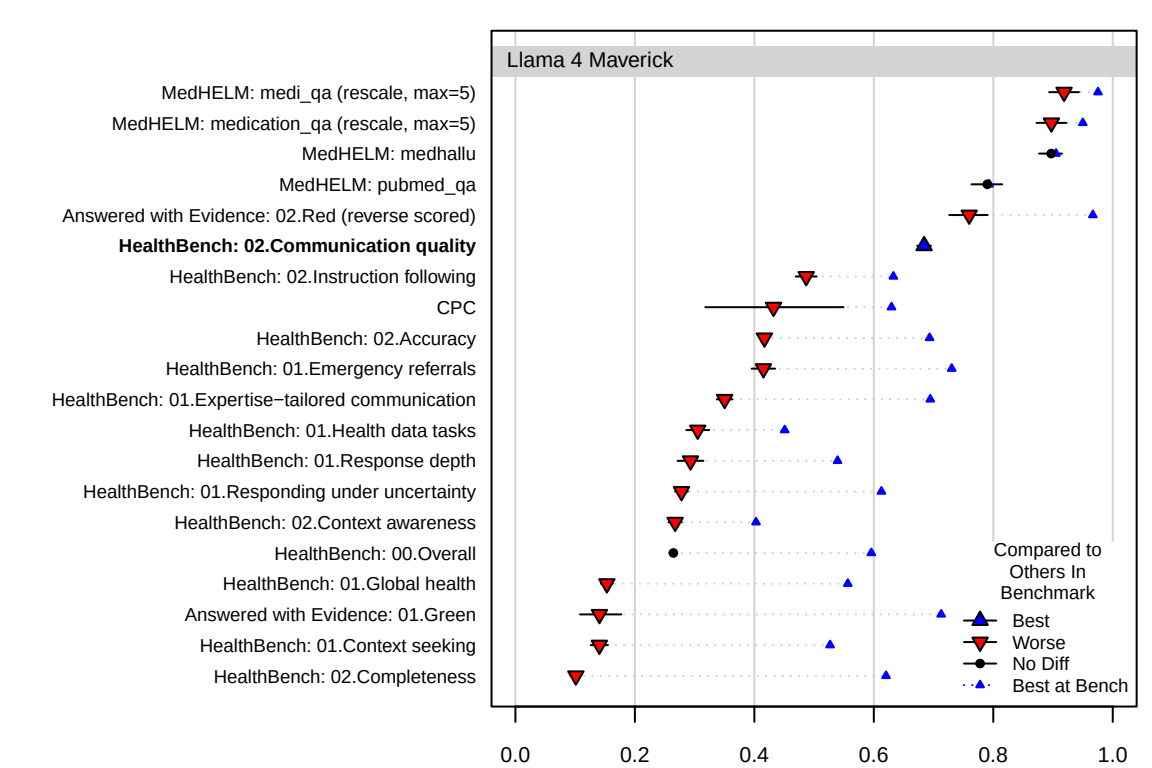
Llama 4 Maverick
MedHELM: medi_qa (rescale, max=5)
MedHELM: medication_qa (rescale, max=5)
MedHELM: medhallu
MedHELM: pubmed_qa
Answered with Evidence: 02.Red (reverse scored)
HealthBench: 02.Communication quality
HealthBench: 02.Instruction following
CPC
HealthBench: 02.Accuracy
HealthBench: 01.Emergency referrals
HealthBench: 01.Expertise−tailored communication
HealthBench: 01.Health data tasks
HealthBench: 01.Response depth
HealthBench: 01.Responding under uncertainty
HealthBench: 02.Context awareness
HealthBench: 00.Overall
HealthBench: 01.Global health
Answered with Evidence: 01.Green
HealthBench: 01.Context seeking
HealthBench: 02.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