

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: Axis: Communication Quality

- HealthBench: Axis: Instruction Following
- CPC
- HealthBench: Axis: Accuracy
- HealthBench: Theme: Emergency Referrals
- HealthBench: Theme: Communication
- HealthBench: Theme: Health Data Tasks
- HealthBench: Theme: Complex Responses
- HealthBench: Theme: Hedging
- HealthBench: Axis: Context Awareness
- HealthBench: .Overall
- HealthBench: Theme: Global Health
- Answered with Evidence: 01.Green
- HealthBench: Theme: Context Seeking
- HealthBench: Axis: Completeness

