Thank you

Key Takeaways

✓ Medical benchmarks: MedQA, USMLE, PubMedQA, MedMCQA, MMLU

✓ Metrics: Accuracy, factuality, consistency, calibration, safety

✓ Human evaluation: Expert protocols, inter-rater agreement, clinical relevance

✓ Continuous monitoring: Dashboards, drift detection, reporting standards

📚 Resources: Papers with Code | Medical AI Benchmarks | TRIPOD Guidelines