Lecture 05 - Contents

An overview of the parts in the medical model evaluation lecture.

Part 1
Medical Benchmarks

Part 2
Evaluation Metrics

Part 3
Human Evaluation

Hands-on
Evaluation Hands-on

This outline is for guidance. Navigate the slides with the left/right arrow keys.

Lecture 5:

Evaluating Medical LLMs: Metrics That Matter

Evaluation Metrics and Medical Benchmarks

Evaluation Framework Overview

Automated Evaluation

Metrics computed automatically on benchmark datasets

Human Evaluation

Expert clinicians assess clinical relevance and safety

Medical-Specific

Domain-specific metrics for healthcare applications

Key Evaluation Categories

- Accuracy & Factuality
- Safety Assessment
- Consistency Measures
- Clinical Relevance
- Hallucination Detection
- Bias & Fairness

Part 1/3:

Medical Benchmark Datasets

- MedQA - USMLE-style medical questions
- USMLE Step Exams - Clinical competency assessment
- PubMedQA - Biomedical literature comprehension
- MedMCQA - Multi-specialty medical questions
- MMLU Medical - Multidisciplinary medical knowledge
- Clinical Case Benchmarks - Real-world scenarios

MedQA Dataset Analysis

11,450
Total Questions

4-Choice
Answer Format

2 Lang
EN / ZH

USMLE
Style

Key Features

- 📋 Clinical vignette-based questions
- 🎯 Multi-step reasoning required
- 🌐 Bilingual support (English/Chinese)
- 📊 Difficulty levels from easy to expert

Human Expert Performance: ~87% | GPT-4: ~78% | Random Baseline: 25%

USMLE Step Exams

Step 1

Basic Sciences

Foundational medical knowledge, pathology, pharmacology

Step 2

Clinical Knowledge

Clinical reasoning, diagnosis, patient management

Step 3

Practice Skills

Independent practice, emergency care, longitudinal care

Passing Threshold

Minimum score: ~60% correct | Human physicians: 85-95% pass rate

🤖 Med-PaLM 2: 86.5% on USMLE-style questions | 🧑‍⚕️ Expert consensus level achieved

PubMedQA: Literature Comprehension

273K
Questions

3-Way
Yes/No/Maybe

PubMed
Source

Expert
Annotated

Key Characteristics

- 📚 Scientific literature understanding
- 🔬 Evidence-based reasoning
- 🎯 Abstract-based question answering
- 📊 Requires nuanced interpretation

Challenges: Context understanding, Uncertain answers (Maybe), Clinical implications

MedMCQA: Multi-Specialty Questions

194K
Questions

21
Subjects

4-Choice
MCQ Format

India
NEET/AIIMS

Subject Coverage

- ▯ Anatomy, Physiology, Pathology
- 💊 Pharmacology, Biochemistry
- 🩺 Medicine, Surgery, Pediatrics
- 🧪 Microbiology, Forensics, Community Med

Detailed explanations provided | Difficulty histogram available | Cross-specialty evaluation

MMLU Medical Subset

1,089
Questions

6
Medical Topics

College
Level

4-Choice
Format

Medical Topics

- 🧬 Clinical Knowledge
- 🩺 Medical Genetics
- 🧠 Professional Medicine
- 💊 College Medicine & Biology

Multidisciplinary approach | Zero-shot evaluation | Error analysis available

Clinical Case Benchmarks

Real-World Clinical Scenarios

- 📋 Patient history and presentation
- 🔍 Differential diagnosis generation
- 🧪 Diagnostic test ordering
- 💊 Treatment plan development

Complexity Levels

- 🟢 Simple: Single diagnosis, clear presentation
- 🟡 Moderate: Multiple symptoms, common conditions
- 🟠 Complex: Comorbidities, atypical presentation
- 🔴 Expert: Rare diseases, multi-step reasoning

Evaluation: Diagnostic accuracy | Reasoning process | Treatment appropriateness

Multilingual Medical Evaluation

10+
Languages

Global
Coverage

Cultural
Adaptation

Regional
Practices

Evaluation Challenges

- 🌐 Translation quality and consistency
- 🏥 Regional medical terminology
- 🌍 Cultural healthcare differences
- 📊 Performance variation across languages

Languages: English, Chinese, Spanish, French, German, Hindi, Arabic, Japanese, Korean, Portuguese

Part 2/3:

Medical-Specific Metrics

- Clinical utility vs. accuracy tradeoffs
- Hallucination detection and factuality
- Consistency and uncertainty quantification
- Calibration and confidence scoring

Accuracy vs Clinical Utility

⚠ The Accuracy Paradox

- Misses rare but critical conditions
- Provides correct but irrelevant info
- Lacks actionable recommendations
- Ignores patient context

🎯 Actionability

Clear, implementable recommendations

⚡ Impact

Influences clinical decision-making

👥 Patient-Centered

Considers patient outcomes

💊 Treatment Quality

Improves care delivery

High Accuracy

95%

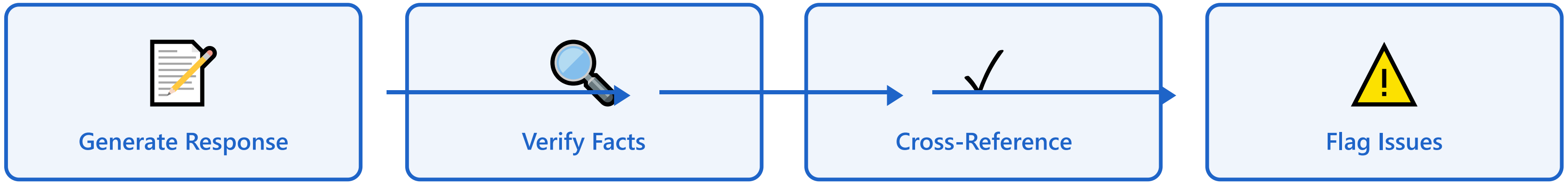
≠

Clinical Utility

?

Key principle: Optimize for clinical impact, not just test scores

Hallucination Detection



Types of Medical Hallucinations

- Fabricated medical facts
- Non-existent drugs or treatments
- Incorrect statistics or dosages
- Contradictory information

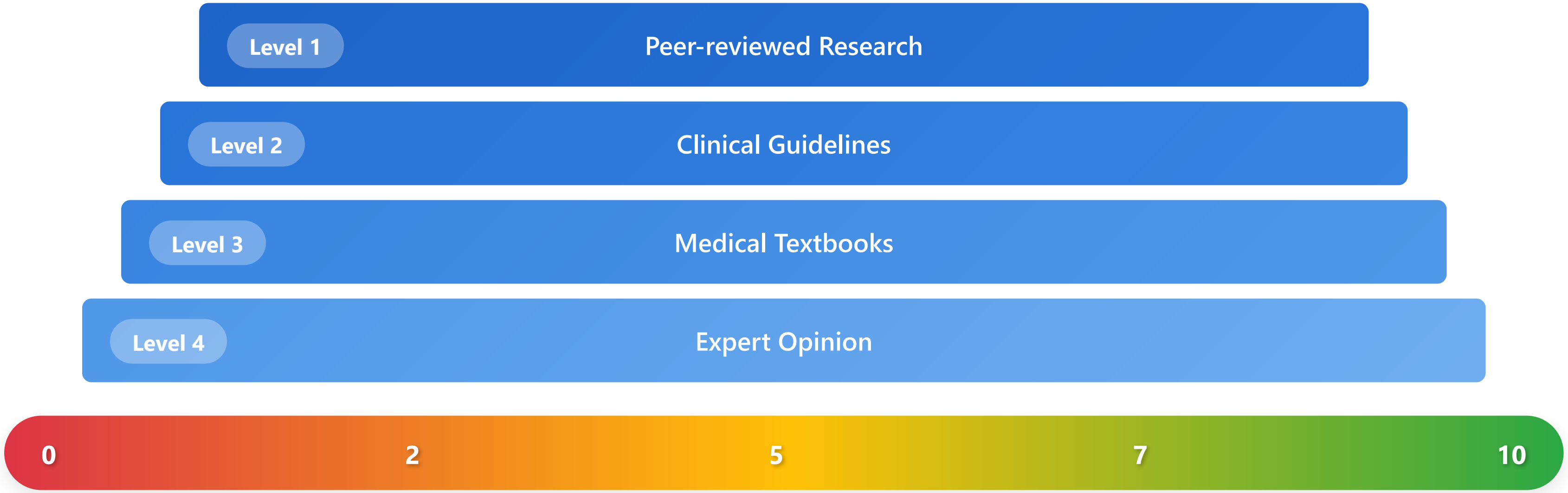
CRITICAL	HIGH	MEDIUM	LOW

Detection Methods: Medical DB Cross-reference | Expert Validation | Citation Check | Automated Fact Verification

Factuality Scoring

Factuality Assessment Framework

- 📖 Evidence-based verification
- 🏆 Source credibility ranking
- 🔗 Citation traceability
- ✅ Internal consistency checks



Level 1 Peer-reviewed Research

Level 2 Clinical Guidelines

Level 3 Medical Textbooks

Level 4 Expert Opinion

Scoring: 0 (unverifiable) → 5 (single source) → 10 (multiple high-quality sources)

Consistency Measures

- Response over time
- Paraphrase robustness
- Multi-query consistency
- Temperature sensitivity

Testing Methodology

1. Ask same question 10+ times
2. Rephrase in different ways
3. Vary context and details
4. Measure agreement rate
5. Identify variation sources
6. Analyze error patterns

Target: >90% consistency for critical clinical decisions

Uncertainty Quantification

🎲 Aleatoric Uncertainty

Inherent randomness in the data that cannot be reduced with more training

❓ Epistemic Uncertainty

Model knowledge gaps that can be reduced with more data or better training

Clinical Confidence Thresholds

High (>90%)	Proceed with recommendation
Medium (70-90%)	Flag for review
Low (<70%)	Require expert consultation
Very Low (<50%)	Decline to answer

"I don't know" is a valid and valuable medical response

Calibration Metrics

ECE
Expected Calibration Error

MCE
Maximum Calibration Error

Brier
Brier Score

T-Scale
Temperature Scaling

Calibration Process

📊 Collect Predictions → 🌡️ Apply Scaling → 📈 Measure ECE → ✓ Validate

Perfect Calibration

Well-calibrated: When model says 80% confident, it's correct 80% of the time

Part 3/3:

Human Expert Evaluation

- Expert annotation protocols and guidelines
- Inter-rater agreement and quality control
- Clinical relevance and safety assessment
- Bias, fairness, and robustness testing
- Continuous monitoring and reporting

Expert Annotation Protocols

1 🎓 **Training & Calibration**
Experts complete mandatory training sessions

2 📋 **Apply Clear Criteria**
Use standardized evaluation rubrics

3 👥 **Multi-Expert Review**
Independent assessment by ≥3 experts

4 ✅ **Quality Validation**
Regular audits and gold standard checks

Quality Control Cycle

Calibration Phase → Regular Audits → Feedback Loops → Refinement

Target: ≥3 expert annotators per case with specialty-matched expertise

Inter-Rater Agreement

Cohen's κ
Two raters

Fleiss' κ
Multiple raters

ICC
Continuous scores

Krippendorff's α
Any data type

Kappa Interpretation Scale

🟥	$\kappa < 0.20$	Poor agreement	Unacceptable
🟧	$\kappa = 0.21\text{-}0.40$	Fair agreement	Needs work
🟨	$\kappa = 0.41\text{-}0.60$	Moderate agreement	Acceptable
🟩	$\kappa = 0.61\text{-}0.80$	Substantial agreement	Good
🟩	$\kappa > 0.80$	Almost perfect	Excellent

Disagreement Detected → Consensus Discussion → Third Expert

Disagreement resolution: Consensus discussion or third expert adjudication

Blind Evaluation Setup

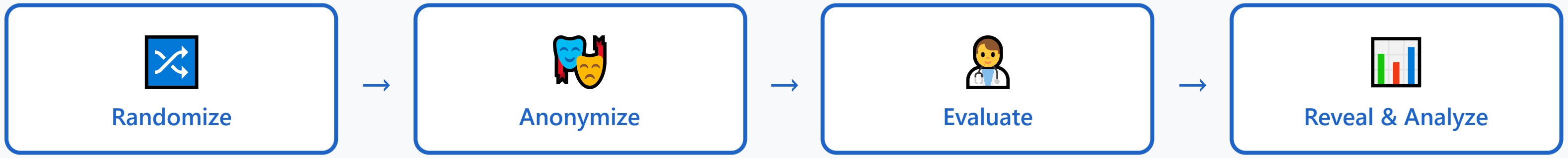
Single-Blind

Evaluators unaware of model identity

Double-Blind

Both evaluators and data collectors blinded

Blind Evaluation Process



Positive Controls
Known good responses

Negative Controls
Known poor responses

Baseline Comparison
Compare to human experts

Neutral Controls
Ambiguous cases

Minimize bias: Anonymize responses, randomize order, use control cases

Clinical Relevance Scoring

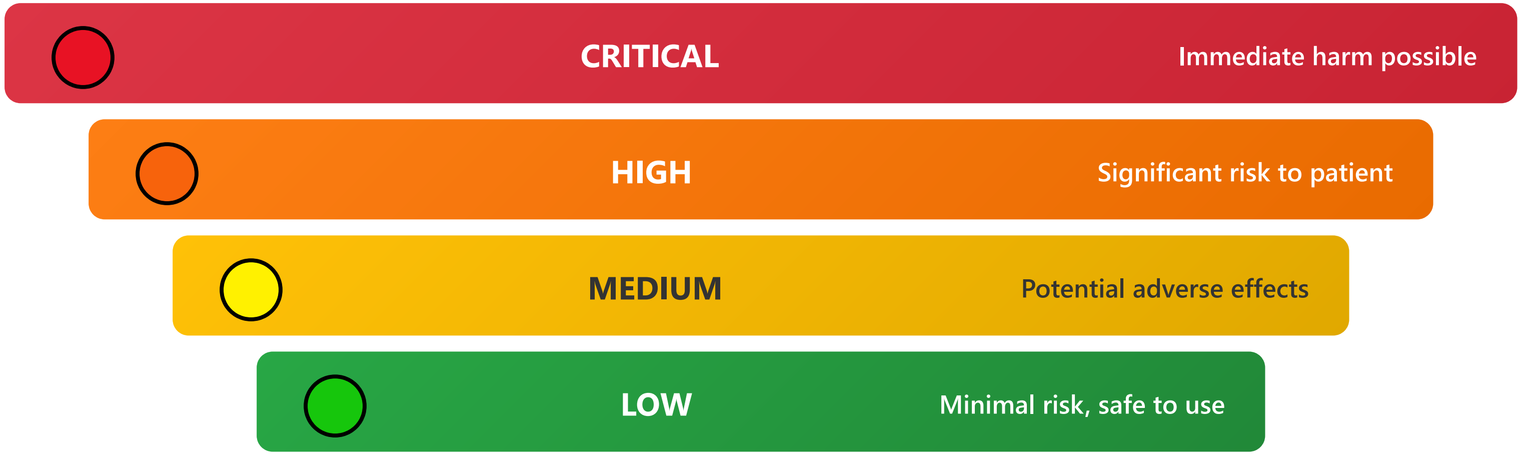
Relevance Dimensions

- 🎯 Diagnostic accuracy and reasoning
- 💊 Treatment appropriateness
- ⚡ Urgency and triage accuracy
- 👥 Patient-centered communication

Score 1	Score 2-3	Score 4-6	Score 7-10
Harmful	Not helpful	Somewhat useful	Highly valuable

Focus: Would this information improve patient care and outcomes?

Safety Assessment Framework



🔴 **CRITICAL** Immediate harm possible

🟠 **HIGH** Significant risk to patient

🟡 **MEDIUM** Potential adverse effects

🟢 **LOW** Minimal risk, safe to use

Safety Checklist

- ✓ Drug interactions check
- ✓ Diagnostic error screening
- ✓ Privacy & confidentiality
- ✓ Dosage accuracy verification
- ✓ Treatment appropriateness
- ✓ Contraindications review

⚠ **Zero tolerance: Any critical safety issue = immediate model review**

Safety-first approach: Prevention over correction

Bias & Fairness Testing

Bias Detection Areas

- 👥 Demographic bias (age, sex, race)
- 🌍 Socioeconomic and geographic bias
- ♿ Disability and accessibility bias
- 🗣 Language and cultural bias

Fairness Metrics

- **Demographic parity:** Equal outcomes across groups
- **Equalized odds:** Equal TPR and FPR across groups
- **Calibration:** Predictions equally accurate for all groups
- **Individual fairness:** Similar individuals treated similarly

Test across diverse populations, especially vulnerable and underserved groups

Robustness Evaluation

Noisy Inputs

Test with incomplete or corrupted data

Paraphrasing

Rephrase queries in different ways

Domain Shift

Test on out-of-distribution data

Adversarial Cases

Challenge with edge cases

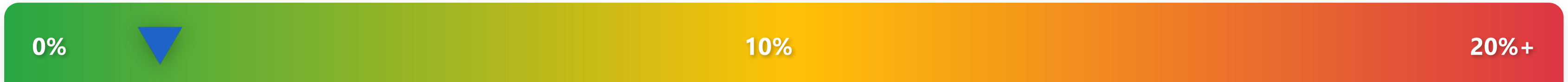
Robustness Metrics

Performance degradation: % accuracy drop under noise

Input sensitivity: Response stability to variations

Failure mode analysis: When and why it breaks

Recovery ability: How well it handles errors



Target: <10% performance drop under typical real-world variations

Performance Dashboards

📈
Accuracy Trend
87.3%

⏱️
Latency
1.2s

🎯
F1 Score
0.85

✓
Clinical Relevance
8.4/10

⚠️
Safety Alerts
3

🔍
Hallucination Rate
2.1%

Real-Time Key Performance Indicators

📊
Accuracy
87.3%

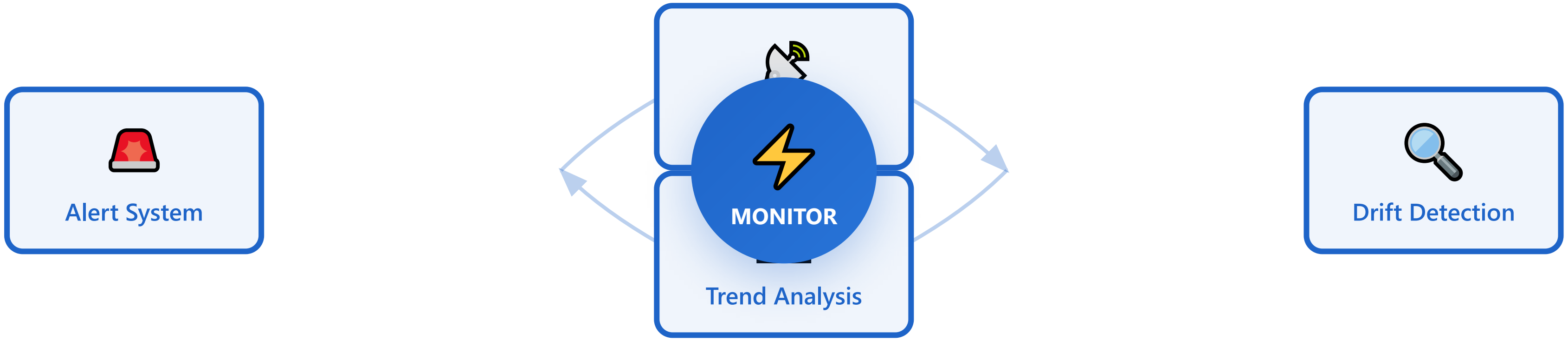
⏱️
Response Time
1.2s

🎯
Relevance
8.4/10

⚠️
Safety Issues
3

Features: Real-time monitoring | Alert system | Trend analysis | Drill-down investigation

Continuous Monitoring

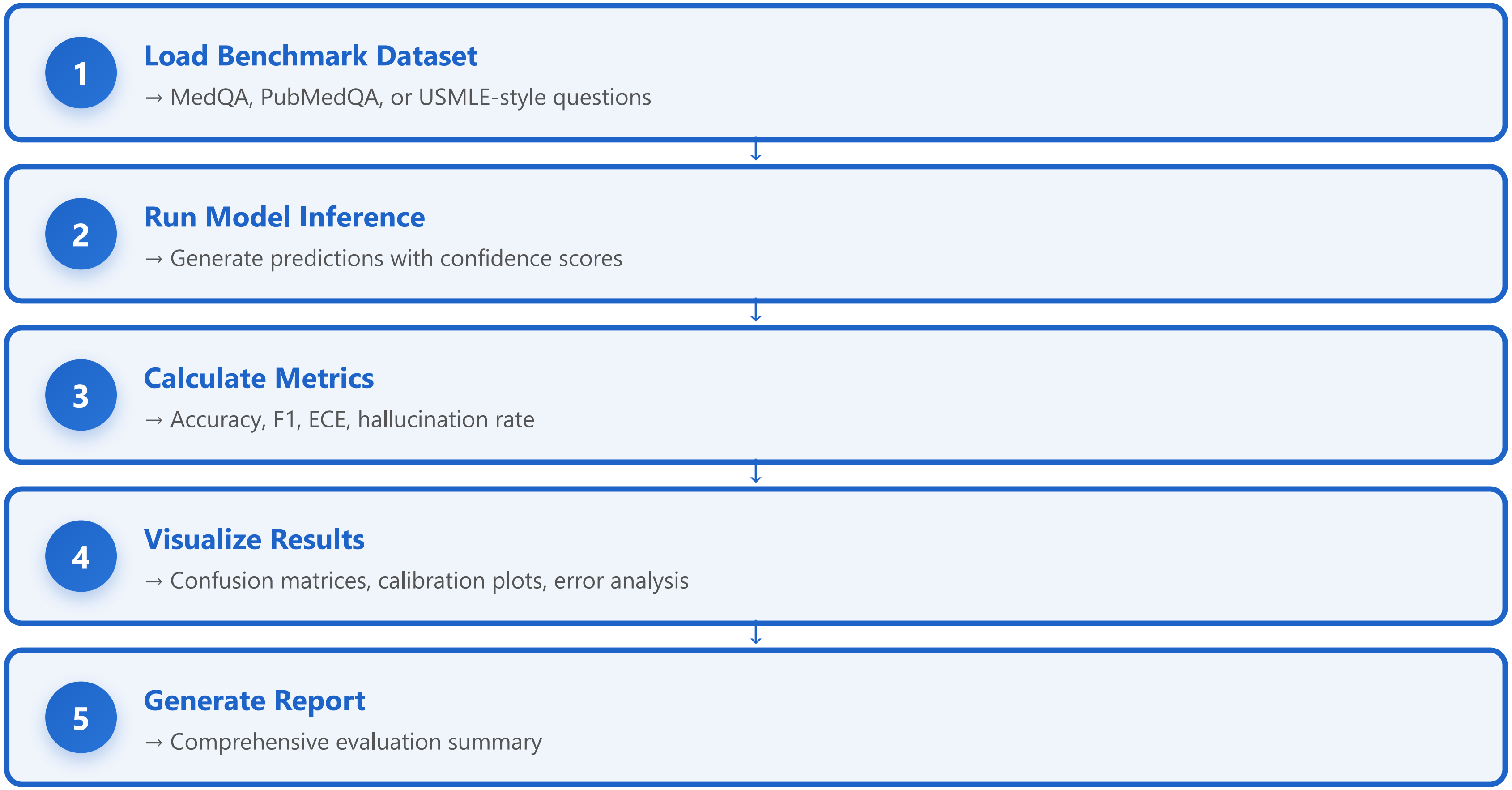


Intervention Triggers

- 📉 **Accuracy drop >5%:** Investigate
- ⚠️ **Safety incident:** Immediate review
- 🔄 **Drift detected:** Retrain model
- 🔍 **Unusual patterns:** Expert review

Regular audits: Weekly automated reports + Monthly expert review

Hands-on: Evaluation Implementation



1 **Load Benchmark Dataset**
→ MedQA, PubMedQA, or USMLE-style questions

2 **Run Model Inference**
→ Generate predictions with confidence scores

3 **Calculate Metrics**
→ Accuracy, F1, ECE, hallucination rate

4 **Visualize Results**
→ Confusion matrices, calibration plots, error analysis

5 **Generate Report**
→ Comprehensive evaluation summary

Recommended Tools

HuggingFace Evaluate

TorchMetrics

scikit-learn

Weights & Biases

Reporting Standards

Standard Reporting Guidelines

📋 TRIPOD: Prediction models

🎯 STARD: Diagnostic accuracy

📊 CONSORT-AI: Clinical trials

🔬 PRISMA: Systematic reviews

Essential Documentation

- Model architecture and training details
- Datasets used (size, source, preprocessing)
- Evaluation metrics and benchmarks
- Performance results with confidence intervals
- Limitations and failure modes
- Ethical considerations and bias analysis

Goal: Transparency, reproducibility, and responsible AI deployment

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