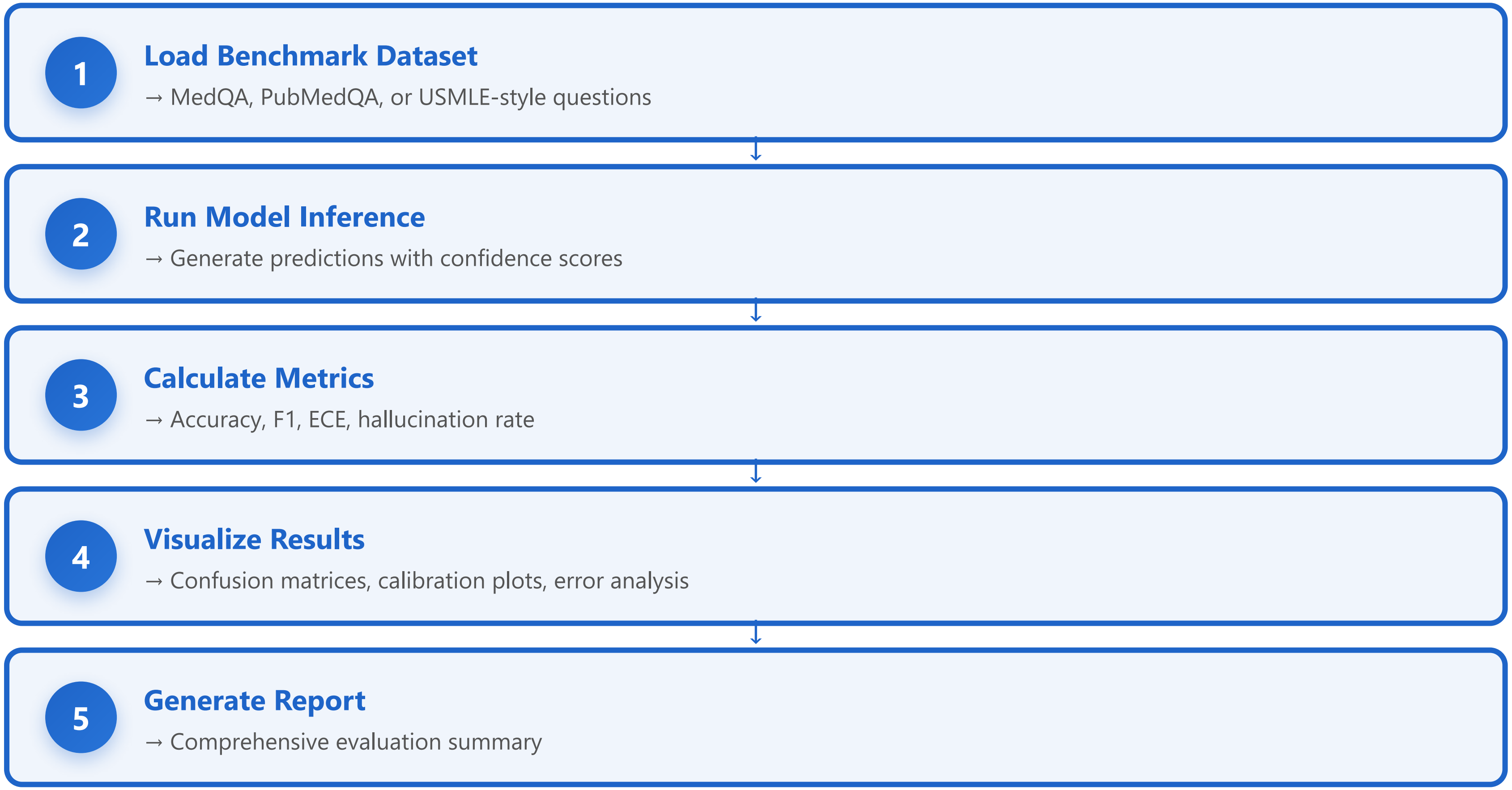
Hands-on: Evaluation Implementation



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