# Document Metadata

author: jfraze@mycomp.org

title: Validation and Testing for Generative AI Systems

nist\_function: Measure

priority\_phase: Must

last\_reviewed: 2025-06-24

status: draft

## **Purpose**

To establish standardized validation and testing protocols for generative AI systems used in clinical, administrative, or operational settings. This ensures alignment with safety, performance, and compliance requirements, including NIST AI RMF, AI 600-1, and SP 800-218A secure development principles.

## **Scope**

This policy applies to all GenAI tools—whether developed in-house or procured—prior to deployment or integration into any hospital-facing system. This includes large language models (LLMs), fine-tuned models, plug-in enabled assistants, and API-connected decision tools.

## **Policy Statement**

All generative AI models must undergo formal validation and testing to verify:  
- Factual accuracy and alignment with domain knowledge,  
- Resistance to adversarial inputs (e.g., prompt injection, jailbreaks),  
- Absence of protected health information (PHI) leakage or unconsented data use,  
- Conformance to stated intended use cases and risk tolerance levels.

## **Roles and Responsibilities**

• Model Validation Team: Executes technical evaluations (e.g., hallucination rate, truthfulness benchmarks, robustness tests).

• Clinical QA Staff: Validates medical relevance and alignment for clinical GenAI.

• Security Analysts: Run adversarial robustness checks and penetration tests on exposed APIs and prompts.

• Legal/Compliance: Reviews validation documentation for FDA, HIPAA, or ONC reporting needs.

## **Implementation Phases**

### **Must Do**

• Implement automated hallucination and truthfulness tests (NIST AI 600-1 §2.3.1).

• Run prompt injection red-teaming and jailbreak simulations (SP 800-218A §5.1.1).

• Establish baseline performance thresholds per intended domain (clinical, operational).

• Log model outputs under controlled conditions for evaluation reproducibility.

• Validate all models against safety-critical failure modes (e.g., treatment recommendation errors).

### **Should Do**

• Conduct fairness testing across relevant demographic slices for model bias exposure.

• Test fallback procedures when the model refuses unsafe or ambiguous requests.

• Include third-party evaluation tools or checklists (e.g., MedQA, TruthfulQA, LLM-as-a-Judge).

• Evaluate consistency of behavior across model versions or prompt variants.

### **Recommended**

• Integrate hallucination KPIs into DevOps pipeline monitoring.

• Use formal verification tools for controlled generation (e.g., constraints on structured outputs).

• Establish peer review of test results before deployment into production systems.

## **References**

• NIST SP 800-218A §5 (Testing and Evaluation)

• NIST AI RMF 1.0: MEASURE-1.3, MEASURE-2.4, MEASURE-2.9

• NIST AI 600-1: §2.3.1, §2.4.1

• Top Vulnerabilities: #1 Prompt Injection, #11 Hallucination, #18 Model Misuse

## **Review Cycle**

• Review prior to each model version deployment

• Re-test if substantial changes to prompts, plug-ins, or third-party dependencies are introduced