# Document Metadata

author: jfraze@mycomp.org

title: Glossary of AI and Generative AI Terms

section: Appendix

last\_reviewed: 2025-06-24

status: draft

# **(SAMPLE)**

# **X1\_glossary.md – AI and GenAI Risk Policy Glossary**

This glossary defines key terms used across the hospital’s AI Risk Management Policy. Terms are based on NIST AI RMF, AI 600-1, HIPAA, and FDA AI guidance, with clinical context where applicable.

## **A**

• AI System  
A machine-based system capable of producing outputs (predictions, recommendations, decisions) given human-defined objectives.

AI System  
A machine-based system capable of producing outputs (predictions, recommendations, decisions) given human-defined objectives.

• Algorithmic Risk  
Potential for errors, bias, or unintended harm due to an algorithm’s design, training data, or implementation.

Algorithmic Risk  
Potential for errors, bias, or unintended harm due to an algorithm’s design, training data, or implementation.

• Accountability  
Requirement for human actors to remain responsible for outcomes even when AI assists or automates decision-making.

Accountability  
Requirement for human actors to remain responsible for outcomes even when AI assists or automates decision-making.

• Adversarial Prompt  
Input crafted to manipulate GenAI outputs (e.g., jailbreak, data leakage), often used in prompt injection attacks.

Adversarial Prompt  
Input crafted to manipulate GenAI outputs (e.g., jailbreak, data leakage), often used in prompt injection attacks.

• Annotation Drift  
Inconsistency in human-labeled data over time or across annotators that can affect supervised model training quality.

Annotation Drift  
Inconsistency in human-labeled data over time or across annotators that can affect supervised model training quality.

• Automation Bias  
Over-reliance on AI-generated outputs, especially in high-stakes or time-constrained environments.

Automation Bias  
Over-reliance on AI-generated outputs, especially in high-stakes or time-constrained environments.

• Auditability  
Capability of an AI system to be independently examined, including traceable logs and documented design intent.

Auditability  
Capability of an AI system to be independently examined, including traceable logs and documented design intent.

## **B**

• Bias (in AI)  
Systematic skew in outputs due to model training data, architecture, or unintended use.

Bias (in AI)  
Systematic skew in outputs due to model training data, architecture, or unintended use.

• Black Box Model  
AI model whose internal logic is not interpretable, especially problematic in clinical or regulated contexts.

Black Box Model  
AI model whose internal logic is not interpretable, especially problematic in clinical or regulated contexts.

• Boundary Violation  
Output that exceeds authorized behavior (e.g., unsolicited diagnosis by a GenAI documentation assistant).

Boundary Violation  
Output that exceeds authorized behavior (e.g., unsolicited diagnosis by a GenAI documentation assistant).

• Bootstrapping  
Use of model-generated data as training input, potentially introducing compounding errors or hallucinations.

Bootstrapping  
Use of model-generated data as training input, potentially introducing compounding errors or hallucinations.

• Behavioral Monitoring  
Real-time or post-hoc surveillance of model outputs to detect anomalies, hallucinations, or misuse.

Behavioral Monitoring  
Real-time or post-hoc surveillance of model outputs to detect anomalies, hallucinations, or misuse.

## **C**

• Clinical Decision Support (CDS)  
System designed to assist clinicians with evidence-based recommendations; regulated if influencing patient care.

Clinical Decision Support (CDS)  
System designed to assist clinicians with evidence-based recommendations; regulated if influencing patient care.

• Content Moderation (LLM)  
Techniques used to filter, redact, or constrain harmful, unsafe, or non-compliant outputs from GenAI models.

Content Moderation (LLM)  
Techniques used to filter, redact, or constrain harmful, unsafe, or non-compliant outputs from GenAI models.

• Corpus  
The body of text or data used to train or fine-tune a GenAI system.

Corpus  
The body of text or data used to train or fine-tune a GenAI system.

• Concept Drift  
Change in statistical properties or semantic meaning in inputs over time, causing model degradation.

Concept Drift  
Change in statistical properties or semantic meaning in inputs over time, causing model degradation.

• Compliance Risk  
Likelihood that an AI system may violate legal, regulatory, or internal policy requirements.

Compliance Risk  
Likelihood that an AI system may violate legal, regulatory, or internal policy requirements.

## **D**

• Data Provenance  
Documentation of data’s origin, history, and transformations across model training and evaluation workflows.

Data Provenance  
Documentation of data’s origin, history, and transformations across model training and evaluation workflows.

• Data Minimization  
Principle of collecting only the necessary data to perform a function, aligned with privacy and security frameworks.

Data Minimization  
Principle of collecting only the necessary data to perform a function, aligned with privacy and security frameworks.

• Drift Detection  
Methods for identifying performance degradation in AI models due to changing input distributions or behaviors.

Drift Detection  
Methods for identifying performance degradation in AI models due to changing input distributions or behaviors.

• Data Leakage  
Unintentional exposure of sensitive or protected data, often due to inadequate input controls or prompt engineering.

Data Leakage  
Unintentional exposure of sensitive or protected data, often due to inadequate input controls or prompt engineering.

• Decommissioning  
Safe and auditable retirement of AI/GenAI systems, ensuring that no residual data or access paths remain active.

Decommissioning  
Safe and auditable retirement of AI/GenAI systems, ensuring that no residual data or access paths remain active.

## **G**

• Generative AI (GenAI)  
AI capable of producing novel content such as clinical notes, summaries, dialogue, or structured text.

Generative AI (GenAI)  
AI capable of producing novel content such as clinical notes, summaries, dialogue, or structured text.

• Governance (AI)  
Oversight frameworks involving policies, stakeholders, and escalation paths for managing AI systems.

Governance (AI)  
Oversight frameworks involving policies, stakeholders, and escalation paths for managing AI systems.

• Ground Truth  
Verified and reliable data used for validating AI model predictions or outputs.

Ground Truth  
Verified and reliable data used for validating AI model predictions or outputs.

• Gating Function  
Controls that limit GenAI functionality or require human confirmation before action (e.g., prior to EHR save).

Gating Function  
Controls that limit GenAI functionality or require human confirmation before action (e.g., prior to EHR save).

• Guideline Conformity  
Degree to which AI outputs align with clinical practice guidelines, medical coding standards, or regulatory norms.

Guideline Conformity  
Degree to which AI outputs align with clinical practice guidelines, medical coding standards, or regulatory norms.

## **H**

• Hallucination (GenAI)  
Output that is plausible-sounding but false, fabricated, or unverifiable—an elevated risk in clinical domains.

Hallucination (GenAI)  
Output that is plausible-sounding but false, fabricated, or unverifiable—an elevated risk in clinical domains.

• HIPAA (Health Insurance Portability and Accountability Act)  
U.S. federal law establishing privacy and security rules for PHI in electronic systems, including AI.

HIPAA (Health Insurance Portability and Accountability Act)  
U.S. federal law establishing privacy and security rules for PHI in electronic systems, including AI.

• Human-in-the-Loop (HITL)  
Workflow structure where human actors review, approve, or edit AI-generated outputs before use.

Human-in-the-Loop (HITL)  
Workflow structure where human actors review, approve, or edit AI-generated outputs before use.

• Handoff Logging  
Documentation when control passes from AI to human or between models, for accountability tracking.

Handoff Logging  
Documentation when control passes from AI to human or between models, for accountability tracking.

• Heuristic Filter  
Rules-based post-processing layer used to block unsafe or biased outputs from GenAI systems.

Heuristic Filter  
Rules-based post-processing layer used to block unsafe or biased outputs from GenAI systems.

## **P**

• Prompt Engineering  
Design and refinement of input text to elicit reliable, compliant responses from a GenAI model.

Prompt Engineering  
Design and refinement of input text to elicit reliable, compliant responses from a GenAI model.

• Prompt Injection  
Security vulnerability where adversarial prompts override system instructions or expose private context.

Prompt Injection  
Security vulnerability where adversarial prompts override system instructions or expose private context.

• Provenance Logging  
Real-time recordkeeping of input prompts, model versions, output types, and decision context.

Provenance Logging  
Real-time recordkeeping of input prompts, model versions, output types, and decision context.

• PHI (Protected Health Information)  
Any health-related information that can identify an individual, subject to HIPAA and data protection policies.

PHI (Protected Health Information)  
Any health-related information that can identify an individual, subject to HIPAA and data protection policies.

• Pretraining Corpus  
Large-scale text data used to train foundational models before domain-specific tuning.

Pretraining Corpus  
Large-scale text data used to train foundational models before domain-specific tuning.

## **S**

• Synthetic Data  
Artificially generated datasets used to augment training or testing, particularly where real data is sensitive.

Synthetic Data  
Artificially generated datasets used to augment training or testing, particularly where real data is sensitive.

• Safety Guardrails  
Embedded model constraints, filtering systems, and human override mechanisms to prevent harm.

Safety Guardrails  
Embedded model constraints, filtering systems, and human override mechanisms to prevent harm.

• System Boundary  
Defined limits within which an AI system is authorized to operate or generate content.

System Boundary  
Defined limits within which an AI system is authorized to operate or generate content.

• Semantic Drift  
Deviation in the intended meaning of a prompt or term over time, impacting prompt consistency or model response.

Semantic Drift  
Deviation in the intended meaning of a prompt or term over time, impacting prompt consistency or model response.

• System Card  
Summary documentation of a GenAI system's intended use, risks, safeguards, and compliance dependencies.

System Card  
Summary documentation of a GenAI system's intended use, risks, safeguards, and compliance dependencies.

## **T**

• Traceability  
Capability to trace AI decisions and outputs back to specific inputs, models, and prompt settings.

Traceability  
Capability to trace AI decisions and outputs back to specific inputs, models, and prompt settings.

• Transparency (AI)  
Degree to which stakeholders can understand how an AI system makes decisions or generates content.

Transparency (AI)  
Degree to which stakeholders can understand how an AI system makes decisions or generates content.

• Training Data Governance  
Policies and controls over the sourcing, labeling, and permission status of data used for AI model development.

Training Data Governance  
Policies and controls over the sourcing, labeling, and permission status of data used for AI model development.

• Temperature (LLM)  
Model parameter controlling randomness in outputs; higher temperature increases variability and hallucination risk.

Temperature (LLM)  
Model parameter controlling randomness in outputs; higher temperature increases variability and hallucination risk.

## **Z**

• Zero-shot Prompting  
Asking a model to perform a task without explicit examples or fine-tuning for that specific instruction.

Zero-shot Prompting  
Asking a model to perform a task without explicit examples or fine-tuning for that specific instruction.

• Zero Trust (AI Security)  
Security principle assuming no implicit trust for model components, users, or data—critical for GenAI deployment boundaries.

Zero Trust (AI Security)  
Security principle assuming no implicit trust for model components, users, or data—critical for GenAI deployment boundaries.