Term Definition

AI Care Standard™
for Patient Communication

AI Definitions

February 2026

AI Judge

A secondary AI model that evaluates the safety, accuracy, appropriateness, and quality of another AI's outputs in real-time. Judge models act as automated quality control layers, scoring primary model outputs and triggering alerts or rejections when problems are detected. This approach can reduce hallucinations by up to 80% by catching errors before they reach patients.

Anomaly Detection

Automated systems that identify when AI outputs deviate from expected patterns or thresholds. In patient communication, anomaly detection can catch when messages contain unusual medical advice, fall outside standard deviations for tone or content, or show signs of potential errors—flagging these cases for human review before release.

Audit Trail

A complete, timestamped record of all AI interactions, decisions, and data sources that enables tracking, accountability, quality review, and investigation if issues arise. Audit trails must capture the model version used, inputs received, sources consulted, outputs generated, and any human review or intervention that occurred.

Benchmark Testing

Systematic evaluation of an AI system's performance against standardized criteria, reference datasets, or established accuracy thresholds to verify it meets required standards. Benchmarks might include medical question-answering datasets, readability tests, or institution-specific clinical scenarios designed to detect drift or degradation.

Confidence Threshold

A minimum level of certainty an AI system must have before providing an answer. When confidence falls below this threshold, the system should acknowledge uncertainty, provide caveats, or escalate to a human provider. Confidence scores are calculated internally based on how well the available information matches patterns the model has learned.

Data Normalization

Converting clinical data from various formats and sources into a standardized structure to ensure consistent handling before it reaches AI systems. Healthcare data arrives in many forms—structured EHR fields, free-text notes, lab results—and normalization ensures the AI interprets this information correctly, regardless of its original format.

Ensemble Methods

Using multiple AI models or validation approaches together to improve accuracy and reduce errors. In healthcare, ensemble methods might involve having two different AI models analyze the same clinical information, then cross-checking their outputs for consistency. Disagreements between models trigger automatic review rather than simply averaging results.

Escalation Protocol

A defined process for automatically routing high-risk situations—such as medical emergencies, safety concerns, or complex clinical questions—from AI systems to human healthcare providers for immediate review and intervention. Protocols specify trigger conditions, routing paths, response time requirements, and accountability for acknowledgment.

Fine-Tuning

The process of taking a pre-trained AI model and further training it on specialized, domain-specific data to improve performance for particular tasks. In healthcare AI, fine-tuning might involve training a general language model on anonymized patient communication examples to better understand medical terminology, appropriate tone for sensitive topics, and healthcare-specific conversation patterns. Fine-tuning must occur in controlled environments using validated datasets with explicit documentation of all changes.

Generative AI

AI technology that creates new content—such as text, images, or responses—based on patterns learned from training data. In healthcare, generative AI can draft patient messages, summarize medical information, or answer questions in conversational language. Unlike rule-based systems that follow pre-programmed scripts, generative AI produces original outputs for each interaction.

Governance Framework

The organizational structure, policies, roles, and oversight processes that ensure AI systems are deployed

safely, monitored continuously, and held accountable to clinical and ethical standards. Governance frameworks define approval processes, oversight committees, performance thresholds, incident response procedures, and accountability assignments.

Grounding

The practice of constraining AI outputs to information that exists in verified source documents rather than allowing the model to generate content based purely on training data patterns. Grounded systems restrict responses to evidence-based, institution-approved sources like EHR data, peer-reviewed guidelines, and institutional policies—preventing the AI from "improvising" or speculating about clinical matters.

Guardrails

Technical and operational safeguards built into AI systems to prevent harmful, inaccurate, or inappropriate outputs. Guardrails include content filters, escalation protocols, confidence thresholds, and human oversight mechanisms. In healthcare, guardrails might automatically block AI from discussing certain sensitive topics, flag when confidence is low, or route urgent situations to human providers.

Hallucinations

When an AI system confidently presents false or fabricated information as fact. In healthcare, this occurs when AI invents medical details, clinical findings, or treatment recommendations that don't exist in the patient's actual records—such as reporting test results that were never performed or recommending medications not prescribed. Hallucinations are particularly dangerous in patient communication because the AI's confident tone can make the false information appear credible. Preventing hallucinations requires AI systems to be "grounded" in verified medical records and to acknowledge when information is uncertain or unavailable.

Large Language Model (LLM)

A type of artificial intelligence trained on massive amounts of text data that can understand and generate human-like language. LLMs power chatbots, AI assistants, and automated communication tools used in patient-facing healthcare applications. These models learn patterns from their training data to predict what text should come next in a conversation or summary.

Model Drift

The gradual decline in an AI system's accuracy and reliability over time, occurring when real-world patterns

change from what the model originally learned. In healthcare, drift happens when patient populations shift (new demographics, emerging diseases like COVID-19), medical practices evolve (updated treatment guidelines, new medications), or documentation styles change acrossfacilities. For example, an AI trained on data from Colorado (lowest U.S. obesity rate) would perform poorly in other states with different patient populations. Without regular monitoring and retraining with current data, a "drifted" model may provide outdated recommendations or miss critical patterns—making continuous performance monitoring and periodic model updates essential safety requirements.

Model Training

The process of teaching an AI system by exposing it to large datasets so it learns patterns, relationships, and appropriate responses. In healthcare, training data must be carefully curated to ensure clinical accuracy, geographic diversity, and representation across socio demographic backgrounds. The quality and breadth of training data directly determines the model's ability to perform safely across diverse patient populations.

Model Validation

Testing and verification processes to ensure an AI model performs accurately, safely, and as intended before deployment and throughout its operational life. Validation includes benchmarking against standardized test sets, accuracy measurement, bias testing, and real-world performance monitoring. Healthcare AI should meet predefined accuracy targets (e.g., ≥98% factual accuracy on audited tasks).

Patient-Facing AI

Any AI system that directly interacts with patients, families, or caregivers—including chatbots, patient portals, voice assistants, discharge summarizers, and automated messaging tools. These systems require higher safety standards than internal clinical tools because they affect patient understanding, decision-making, and wellbeing.

Prompt Engineering

The practice of carefully designing the instructions and context provided to an AI system to shape its behavior and outputs. In patient communication, prompt engineering involves specifying requirements like reading level, tone, content to include or exclude, and safety guardrails. Well-designed prompts can significantly improve the accuracy, appropriateness, and safety of AI-generated messages without changing the underlying model.

Provenance

Documentation of where information originated, including the source documents, publication dates, version numbers, and approval signatures that establish an AI's output is based on verified, authoritative medical knowledge. Provenance tracking ensures every clinical statement can be traced back to its evidence base for audit and accountability purposes.

Readability Scoring

Automated assessment of how easy text is to understand, typically measuring grade level, sentence complexity, and medical jargon usage. Healthcare AI should target 6th-8th grade reading levels for general patient communication, with systems automatically flagging content that exceeds readability thresholds for human review or revision.

Retrieval-Augmented Generation (RAG)

A technique that combines AI language generation with real-time information retrieval from verified databases. Instead of relying solely on patterns learned during training, RAG systems first search through approved medical literature, clinical guidelines, or patient records, then use that retrieved information to generate accurate, grounded responses. This approach significantly reduces hallucinations by anchoring AI outputs to authoritative sources rather than allowing "free-styling" based on training data alone.

Retraining

The process of updating an AI model with new data to maintain accuracy over time, incorporate updated medical guidelines, correct errors, or adapt to changing patterns in real-world use. Healthcare AI systems require regular retraining (minimally annually) to prevent drift as medical knowledge evolves, new treatments emerge, and patient populations change.

Sentiment Analysis

AI capability to detect and respond to the emotional tone or distress level in patient communications, enabling appropriate adjustment of response style or triggering escalation when needed. Sentiment analysis helps AI recognize when patients are anxious, confused, or in distress—allowing the system to adapt its communication approach or route the conversation to a human provider.

Shadow Mode

A safety testing approach where a new or updated AI system runs silently alongside the current system, processing the same patient information but without its outputs being sent to patients or affecting care. This allows healthcare teams to compare the new AI's performance against the proven system, catching problems before patients are exposed to potential errors. Shadow mode testing is particularly important when deploying AI at a new hospital facility or after significant updates, ensuring the system performs accurately in that specific clinical environment before releasing any patient-facing communications.

Telemetry

Real-time collection and transmission of performance data from AI systems to monitoring dashboards. Telemetry captures metrics like accuracy rates, escalation frequency, response times, confidence scores, and usage patterns—enabling immediate detection of problems and supporting continuous improvement efforts.

Version Control

Systematic tracking of all changes, updates, and configurations made to an AI system over time, ensuring accountability and enabling rollback if problems emerge. Version control documents which model version, prompt template, data sources, and system settings were active at any given time—critical for investigating incidents and maintaining regulatory compliance.