AI Deep Dive

Quick Reference Guide | SunHawk Consulting

SUNHAWK CONSULTING

THE KEY IDEA: Generative AI is a prediction algorithm; it predicts the most likely next word based on patterns in its training data. Every risk below traces back to this root cause. **Lower temperature = more predictable. Higher temperature = more creative.**

5 RISK CATEGORIES OF USING GENERATIVE AI

Hallucinations

AI generates outputs that sound authoritative but are completely fabricated, and AI doesn't know it's wrong.

Sycophancy

AI is trained on human feedback that rewards agreeable responses. It's trained to tell you what you want to hear, even when AI knows you're wrong.

Prompt Sensitivity

The same question asked by two different users produces different answers due to the probabilistic nature of the model.

Prompt Injection

Bad actors deliberately exploiting AI applications. The model can't distinguish instructions from data.

Drift & Inconsistency

The AI output today may be different tomorrow. Updates to the model or data change probability distributions.

4 TECHNIQUES TO MANAGE THE RISK

Risk Assessment

Classify AI tools by risk tier before deployment. Evaluate what data it touches, who it affects, and what happens when it's wrong.

Human-in-the-Loop

Match oversight to risk: expert validation for clinical decisions, review-before-action for records, spot-checks for low-risk tasks.

Guard Models

A second AI screens inputs/outputs of the primary model, catching injection attempts, hallucinations, and policy violations in real time.

Continuous Monitoring

AI is not set-and-forget. Ongoing monitoring catches drift, vendor changes, and degraded accuracy that a one-time assessment will miss.

KEY RULE: If your AI (A) processes untrusted input, (B) accesses sensitive data, AND (C) takes external actions, human-in-the-loop should be required. (Meta's "Rule of Two")

WHERE YOUR AI LIVES: 3 DEPLOYMENT OPTIONS

Third-Party

- SaaS or API (OpenAI, etc.)
- Data leaves your network
- Fastest to deploy, lowest cost
- Best for: low-risk use cases

Private Cloud

- Azure, AWS, or GCP tenant
- Data stays in your cloud
- Shared responsibility model
- Best for: moderate-risk workloads

On-Premises

- Your own infrastructure
- Data never leaves your walls
- Full organizational control
- Best for: sensitive/high-stakes data

← *Easier to Start* *Easier to Control* →

SUNHAWK CONSULTING

Jacob Mills Jacob.Mills@SunHawkConsulting.com | (859) 421-2514
James Rough Jim@SunHawkConsulting.com | (602) 334-5522

SunHawkConsulting.com

AI Glossary

Key Terms and Definitions

SUNHAWK CONSULTING

FOUNDATIONAL AI TERMS

Algorithm — A set of rules or steps a computer follows to solve a problem or complete a task. An AI model is built on algorithms.

Artificial Intelligence (AI) — A broad term for computer systems designed to perform tasks that normally require human thinking, like recognizing patterns, making decisions, or understanding language.

Chatbot — A software application that uses AI to have text-based conversations with users. Examples include ChatGPT, Claude, and Google's Gemini.

Generative AI (GenAI) — A type of AI that creates new content — text, images, code, or audio — by predicting patterns from data it was trained on. It generates, rather than just retrieves.

Large Language Model (LLM) — The specific type of AI behind tools like ChatGPT and Claude. A massive statistical model trained on text data to predict and generate human language.

Model — The trained AI system itself. Think of it as the "brain" that has learned patterns from data. Different models have different capabilities and risks.

Parameters — Internal settings learned during training that determine how a model makes predictions. More parameters generally mean more capability. GPT-4 has roughly 1.8 trillion.

Prompt — The input you give to an AI — your question, instruction, or request. The quality of the prompt heavily influences the quality of the output.

Token — The unit of text an AI reads and generates. Roughly ¾ of a word. When people say a model accepts "128K tokens," that's how much text it can process at once.

Training Data — The massive collection of text, books, websites, and code used to teach an AI model. The quality of this data shapes what the model knows and how it behaves.

PRESENTATION TERMS

API — Application Programming Interface. A way for software systems to communicate. When a healthcare app "uses AI," it's often sending requests to an AI model through an API.

Autoregressive Generation — The process of generating text one word at a time, where each new word is based on all the words that came before it.

Drift — When AI behavior changes over time due to vendor model updates or changes in the data the AI references, even if no one intentionally changed anything.

Foundation Model — The base AI model produced by pre-training. Powerful but raw and unusable on its own until post-training adds safety and behavioral refinements.

Guard Model — A smaller, purpose-built AI that sits between the user and the primary AI to screen for risky inputs and catch problematic outputs.

Hallucination — When AI generates information that sounds confident and authoritative but is completely fabricated. The AI doesn't know it's wrong.

Human-in-the-Loop (HITL) — A process where a human reviews, validates, or approves AI output before it's acted on. The level of review is matched to the level of risk.

On-Premises — AI infrastructure hosted on your organization's own physical servers. Data never leaves your building. Maximum control, highest cost.

Post-Training (Fine-Tuning) — The second phase of building a model where it's refined for safety, helpfulness, and specific behaviors. Cheaper and faster than pre-training.

Pre-Training — The first and most expensive phase of building an AI model, where it learns language patterns from trillions of words of text data.

Private Cloud — AI hosted in a dedicated cloud environment (Azure, AWS, GCP) controlled by your organization. Data stays in your tenant, but infrastructure is managed by the cloud provider.

Probabilistic Text Generation — How GenAI actually works: it calculates the probability of each possible next word and picks one. It's predicting, not thinking.

Prompt Injection — A deliberate attack where someone hides malicious instructions inside input data to trick the AI into ignoring its safety rules.

Prompt Sensitivity — The tendency for AI to give different answers to the same question depending on how it's phrased or who asks it.

Sycophancy — AI's tendency to agree with you even when you're wrong, because it was trained to give responses people rate as helpful and agreeable.

System Prompt — Hidden instructions set by developers that define how the AI behaves — its role, tone, and rules. Users typically don't see these.

Temperature — A setting that controls how creative or conservative the AI's word choices are. Low = predictable and repeatable. High = more varied but riskier.

Third-Party AI — AI accessed through an external vendor's platform or API. Fastest to deploy, but data leaves your network and you have limited control over the model.

SUNHAWK CONSULTING
Jacob Mills Jacob.Mills@SunHawkConsulting.com | (859) 421-2514
James Rough Jim@SunHawkConsulting.com | (602) 334-5522
SunHawkConsulting.com