AI Security Checklist

Product-level AI security for teams shipping LLM features

Prepared by	Nazline Mwita, CompTIA Security+
Practice	HarLyn Digital Partners
Scope	Self-assessment starting point — not a substitute for a scoped audit

Model & API Access

- AI provider API keys are stored server-side only, never shipped in client bundles.
- Per-feature rate limits and cost ceilings are enforced to prevent abuse-driven billing spikes.
- Model outputs are treated as untrusted input before being used in any downstream action.

Data Exposure

- Prompts sent to third-party model providers are reviewed for accidental PII/secret leakage.
- Vendor data-retention and training-use terms are known and acceptable for the data being sent.
- User-facing AI features have a clear boundary on what data the model can see per request.

Prompt Injection Surface

- User-supplied content that reaches the model (messages, uploaded docs, retrieved web content) is treated as untrusted.
- The system prompt is not the only defense — output is validated before triggering actions.
- Tool-calling / function-calling outputs are schema-validated before execution.

Governance

- There's a named owner for AI-feature risk decisions, not just “whoever shipped it.”
- AI-generated content shown to users is distinguishable from human-authored content where it matters.
- There's an incident path for when an AI feature produces harmful or clearly wrong output.

How to use this checklist

This is a starting-point self-assessment, not a substitute for a scoped engagement. Items checked “no” or “unsure” are candidates for a Secure Digital Workflow Assessment.

Nazline Mwita, CompTIA Security+ certified Cybersecurity Assurance Lead and Co-Founder at HarLyn Digital Partners. Specializing in authorized web & API security assessments, KDPA compliance reviews, and defensive cloud architecture in Nairobi, Kenya.