Apptega

GUIDE

NIST AI 600-1 (Generative AI) Compliance Guide

For organizations that design, develop, deploy, procure,
or use Generative AI (GAI) systems and need a practical path
to implement NIST AI 600-1 for GAI

Current as of August 29, 2025

1. Introduction

This guide converts NIST AI 600-1 (Generative AI) into concrete steps that help organizations become compliant with GAI-specific risks and controls. It follows the same four core functions used across modern AI governance—GOVERN, MAP, MEASURE, MANAGE—and provides owners, parameters, acceptance criteria, and evidence tailored to text, image, audio, code, and multimodal generation.

1A. Beginner Quick-Start (First 30—90 Days)

Days 1—15 — Program stand-up

- Appoint: Executive GAI Risk Owner, Responsible AI Lead, Model/Use-Case Owner(s), Data Owner, TEVV Lead, Security/Privacy Lead, Legal/IP & Ethics Counsel, Third-Party/Procurement Lead, Safety/Red-Team Lead.
- Define the GAI authorization boundary (training/fine-tuning/inference pipelines, vector stores, RAG sources, safety services, UI/tool-use, plug-ins/functions).
- Publish GAI Policy and Risk Appetite; create Use-Case Intake and Review Gates.
- Open a GAI Registry (AIBOM) per model/use case.
- Draft GAI Profile (v0) for the first use case (context, trust priorities, thresholds, risk tier).

Days 16—45: Implement & tailor

- Complete MAP: data rights & provenance, creative domain(s), affected parties, foreseeable harms (e.g., misinformation, privacy leak, IP infringement, unsafe content).
- Stand up MEASURE baselines: metrics for safety/toxicity, hallucination/factuality, robustness to jailbreaks/prompt injection, bias/fairness, privacy/PII leakage, IP/Copyright duplication, uncertainty/calibration; define test suites; plan red-teaming.
- Define MANAGE controls: content filters, tool-use allow/deny lists, function-calling constraints, kill-switch & rollback, watermark/provenance stance, monitoring plan, incident taxonomy, incident taxonomy.

Days 46—90: Validate & evidence

- Execute TEVV (safety & misuse probes, hallucination, privacy/IP leakage tests); conduct red-team for jailbreak/injection; remediate gaps.
- Populate the Evidence Register and run a GAI incident tabletop (egress exfil via function tools; unsafe image generation; code with vulnerable patterns).
- Update GAI Profile (v1); record residual risks and owners.

2. Scope & Alignment

Scope. People, processes, datasets (including synthetic), prompts/system instructions, code, pipelines (training, fine-tuning, RAG, inference), safety layers, tool/function integrations, UIs, telemetry, and third parties that create, receive, maintain, transmit, or can materially affect GAI outputs or their impacts. Includes plugins, API tool-use, vector stores, and downstream content distribution channels.

Key definitions

- **GAI system:** Model + runtime that generates content (text, images, audio, code, multimodal) for users or services.
- **TEVV (GAI):** Testing/Evaluation/Verification/Validation, including red-teaming against unsafe content, jailbreak, PII/IP leakage, hallucination.
- **GAI Profile:** Use-case–specific statement of context, outcomes, thresholds, and constraints; updated at lifecycle gates and material changes.
- **Trustworthiness priorities (GAI):** validity/reliability; safety; security/resilience (incl. prompt/plug-in abuse); accountability/transparency; explainability/traceability; privacy; fairness; IP/copyright & licensing compliance; content provenance.

Roles (minimum). Executive GAI Risk Owner; Responsible AI Lead; Model/Use-Case Owner; TEVV Lead; Security/Privacy Lead; Legal/IP & Ethics; Procurement/TPRM; Safety/Red-Team Lead; SRE/Model Ops.

3. Standards & Practices (Mapped to 600-1 Functions)

For each family: Intent • Minimums • Implement (Procedural/Technical/Contractual) • Evidence • Acceptance Criteria • Common Failures • Internal QA Plan • Documentation tie-in.

GOVERN (cross-cutting)

Intent: Establish accountable, transparent, and ethical governance for GAI.

Minimums: Policy; roles & independence; risk tiering; intake + gates; AIBOM/registry; documentation standards.

Implement—Procedural:

- **Use-Case Intake (must complete):** objective; stakeholder/affected groups; legal/contract drivers; **data rights** basis; **IP licensing posture** (training, fine-tuning, inference); risk tier; monitoring & rollback owners.
- **Pre-Launch Readiness Gate:** current GAI Profile; **data rights/provenance attestation**; TEVV reports vs thresholds; **independent evaluator** sign-off (per tier); guardrails configured & tested; **RAG/source governance** verified; monitoring + rollback **proven in a drill; AIBOM entry complete**; residual risks accepted by Executive GAI Risk Owner.
- **Risk Tiering Matrix (3×3 Likelihood×Impact):**

Tier	Example drivers	Independence	Monitoring	Drills
Low	Internal use, reversible, low harm	Internal peer	Monthly	Semiannual
Medium	External or sensitive content	Separate team	Biweekly + bias slices	Quarterly
High	Safety/rights-impacting, broad reach	Independent org + exec sign	Weekly + quarterly red-team	Quarterly + annual full rollback

Implement—Technical: **AIBOM minimum fields** (use-case ID; model/version; **system prompts/policies**; datasets/versions; **RAG sources & retrieval policy**; guardrails; eval reports/metrics; known failure modes; rollout/rollback artifacts; owners; dependencies; license/usage limits).

Implement—Contractual: Provider contracts with evaluation rights, data/usage limits, IP indemnity stance, incident notice timelines, update transparency, watermark/provenance commitments (where available), and **flow-downs**.

- **Evidence:** Policy; role matrix; risk appetite; intake records; gate minutes; **AIBOM entries**; supplier responsibility matrices.
- **Acceptance Criteria:** Each GAI use case has an owner, current profile, documented gates, monitoring plan; **independent evaluation** recorded for medium/high tier.
- **Common Failures:** No registry; vendor claims not independently evaluated; missing RAG/source governance.
- **Internal QA Plan:** Quarterly audit of three launches for gate adherence, AIBOM completeness, and documentation quality.
- **Documentation tie-in:** GAI Policy; Risk Appetite; Review Gate SOP; Model Registry/AIBOM; Supplier Responsibility Matrix.

MAP

Intent: Define context, rights, affected parties, creative constraints, and foreseeable harms.

Minimums: Use-case profile; environment; legal/contract/IP drivers; **data inventory & provenance**; harms; success/exit criteria.

Implement—Procedural:

- **Data Rights Decision Rules** for training/fine-tuning/inference, including **synthetic data**, web crawl, and licensed corpora; **no-train**/derivative limits; retention; geography; redistribution; DSAR/record request handling (as applicable).
- **RAG Source Governance:** allowed repositories; freshness/authority; citation/ grounding rules; cache TTL; disallowed sources; attribution policy.
- **Affected Parties & Harm Scenarios:** direct users, bystanders, subjects of depiction/inference; harms across **safety (harmful content), privacy (PII exposure), IP (copyright/trademark/likeness), misinformation/hallucination, fairness, security** (prompt injection/tool abuse), accessibility. **Assign mitigation owners.**

Implement—Technical: Data lineage; dataset & prompt versioning; content domain constraints (e.g., medical/legal); sandboxing; watermark/provenance labeling plan.

Implement—Contractual: Data/content licenses; training/fine-tuning rights; output usage limits; attribution terms.

- **Evidence:** Use-case profiles; data maps; rights/consent attestations; harm analysis; RAG/source allow-list; environment constraints.
- **Acceptance Criteria:** Profile approved; rights validated; RAG governance documented; harms mapped to mitigations; exit criteria defined.
- **Common Failures:** Missing proof of training/fine-tuning rights; unmanaged RAG sources; undefined depiction/likeness rules.
- **Internal QA Plan:** Review two profiles/quarter for rights, RAG rules, and harm coverage.
- **Documentation tie-in:** GAI Profile; Data Inventory & Lineage; RAG Governance Standard; Harm Analysis SOP.

MEASURE

Intent: Quantify GAI behavior and risks; validate thresholds for safe deployment.

Minimums: **GAI metric suite**; TEVV plans; **independent evaluators** (per tier); reproducibility.

Implement—Procedural: Metric definitions & thresholds; TEVV schedules (pre-launch, periodic, event-driven); versioned eval reports; **GAI red-team** plan (jailbreak/injection, unsafe content, disallowed tools).

Implement—Technical:

- **Safety/Content:** toxicity/violence/sexual content categories; **attack success rate (ASR)** for jailbreak/prompt injection; function-calling abuse rate; code security tests for generated code.
- **Hallucination/Factuality:** grounded answer accuracy; citation adherence for RAG; contradiction rate; unsupported claim rate.
- **Robustness:** perturbation tests; OOD prompts; adversarial prompting; rate-limiting/abuse defenses.
- **Bias/Fairness:** subgroup disparities across defined slices for refusals, unsafe outputs, and helpfulness.
- **Privacy/PII Leakage:** membership inference; PII echo tests; prompt log handling; system prompt exposure tests.
- **IP/Copyright:** near-duplicate detection against training corpora; copyrighted/brand/likeness misuse checks.
- **Uncertainty/Calibration:** refusal/deferral when low-confidence; calibrated thresholds (ECE).

Implement—Contractual: Evaluation rights; telemetry/log access; update disclosures; watermark/provenance verification hooks.

- **Default GAI thresholds (tailor per tier/use case):**
  - **ASR (jailbreak/injection):** ≤ 1% High, ≤ 3% Medium under test battery.
  - **Unsafe content precision/recall:** ≥ 0.95/0.90 for High when filters enabled.
  - **Hallucination (unsupported claims):** ≤ 5% High, ≤ 10% Medium on grounded tasks; RAG citation adherence ≥ 95%.
  - **PII leakage:** 0 confirmed leaks in targeted tests; membership inference not statistically significant ($\alpha=0.05$).
  - **IP duplication:** near-duplicate rate ≤ defined ppm; **blocked classes** (trademark/ likeness) at 100% enforcement in tests.
  - **Bias parity gap:** ≤ 5 percentage points across priority slices or documented justification + mitigation.
  - **ECE (calibration):** ≤ 0.05 High; ≤ 0.08 Medium.
- **Evidence:** TEVV plans; metric dashboards; bias/robustness/privacy/IP reports; red-team results; uncertainty & drift analyses.
- **Acceptance Criteria:** Metrics meet thresholds with limitations documented; known failure modes cataloged; residual risks accepted by Executive GAI Risk Owner; runs are reproducible by an independent party.
- **Common Failures:** Measuring only “helpfulness”; no grounded factuality; missing IP/privacy tests; red-teaming without remediation.
- **Internal QA Plan:** Independently reproduce one evaluation/quarter; spot-check three bias slices; re-run jailbreak suite after each safety update.
- **Documentation tie-in:** TEVV Plan; Metrics Catalog; Red-Team Playbook; Evaluation Reports.

MANAGE

Intent: Treat risks; enforce guardrails; deploy safely; monitor; respond to incidents/change.

Minimums: Risk treatment decisions; **guardrails**; change control; monitoring; **GAI incident taxonomy**.

Implement—Procedural: Risk treatment register; **human-in-the-loop rules** (when review is required; evidence available to reviewers; override logging); **incident taxonomy** (privacy leak, unsafe output, IP infringement, hallucination harming users, security abuse via tools, model/prompt exposure); SLAs & escalation; rollout/rollback & kill-switch drills.

Implement—Technical:

- **Guardrails (pre-launch "must be on"):** input validation; output filters (safety/IP/PII); **RAG grounding & citation enforcement; tool-use allow/deny list**s; function-calling constraints (schema & limits); secrets handling; rate/abuse controls; uncertainty-aware UX (warnings/deferral); **watermark/provenance checks** (where supported); automated drift detection; rollback artifacts.
- **System Prompt & Prompt-Log Security:** restrict visibility; redact PII/secrets; signed versioning; rotation on exposure.

Implement—Contractual: SLA/OLA for quality and safety; incident notice timelines; update transparency; watermark/provenance and attribution commitments (where supported).

- **Evidence:** Risk treatment decisions; guardrail configs & tests; RAG & tool-use policies; monitoring dashboards; incident logs & AARs; rollback test results; change approvals.
- **Acceptance Criteria:** Guardrails active & tested; **citation/grounding enforced**; incidents handled within SLA; material changes re-evaluated; rollback verified.
- **Common Failures:** Tool/function abuse paths open; no provenance stance; logging prompts without PII safeguards; missing kill-switch.
- **Internal QA Plan:** Pre-launch control verification & rollback simulation; quarterly GAI incident drill; periodic watermark/provenance verification.
- **Documentation tie-in:** Risk Treatment Register; Deployment & Rollback SOP; Monitoring Runbooks; Incident Response Plan.

4. GAI Authorization Boundary & Data / Model / Workflow Mapping

Map components, trust zones, datasets & RAG sources, vector stores, system prompts/policies, tool/function integrations, safety layers, telemetry, and external services. Show input ingress (prompts, tools, RAG), output egress (UIs, APIs, publishing), critical safety/privacy/IP control points, and who may change what. Keep under change control.

5. Documentation Set — Canonical Narratives & GAI Profile

Provide narratives that prove conformance to GOVERN/MAP/MEASURE/MANAGE for each scoped GAI use case: mission; roles & accountability; training/fine-tuning rights; RAG/source governance; safety & IP posture; TEVV approach; guardrails; monitoring; incident/change handling; dependencies. Maintain a GAI Profile per use case and keep it current at lifecycle gates and material changes.

6. Applicability & Acceptance Criteria Mapping (to 600-1 Functions)

Function	Outcome Area	Applicable	Implementation Summary	Parameters	Acceptance Criteria	Inheritance	Evidence
GOVERN	Ownership & Gates	Y	Policy; role matrix; risk tier; pre-launch gate	Gate set; tiers	Gate minutes; registry entry	Some vendor	Policy/Minutes
MAP	Rights & RAG Governance	Y	Rights attestation; source allow-list; harm mapping	RAG TTL; rights	Profile approved; rights validated; harms→owners	None	Profiles/Maps
MEASURE	Safety/ Hallucination/ Privacy/IP	Y	Metric suite; red-team; leakage & duplication tests	Thresholds	Meets thresholds; residual risk accepted	None	Eval Reports
MANAGE	Guardrails/ Monitoring/IR	Y	Filters; tool lists; provenance; IR	SLAs	MTTR ≤ SLA; rollback tested	Some vendor	Runbooks/Logs

7. Program Parameters (Organization-Defined Settings for GAI)

- **Governance:** Gates (design, pre-launch, post-launch); independence per tier; documentation minimums.
- **Identity & Change:** Who may modify system prompts, guardrails, tool lists; approvals & audit trails; signed configs.
- **Data & Rights:** Allowed sources; licenses; no-train lists; retention; geography; DSAR handling (as applicable).
- **RAG:** Source allow-list; grounding rules; citation policies; cache TTL; disallowed domains.
- **Security & Privacy:** Threat model for prompt/tool abuse; secret handling; PII scrubbing; prompt log retention/redaction.
- **Trust Targets (defaults):** ASR, hallucination, bias, privacy/IP, calibration thresholds (see **MEASURE**).
- **Monitoring SLOs:** Drift/bias alert MTTD ≤ 24h (High) / ≤ 72h (Medium); incident MTTR ≤ 24h (High) / ≤ 72h (Medium).
- **Re-evaluation triggers:** Model/data/prompt/guardrail/tool or dependency change; threshold breach; adverse incident.
- **Third-Party:** Evaluation rights; incident notice ≤ 72h; update transparency; watermark/provenance stance; flow-downs.

8. Evidence Register

Artifact	Function(s)	Location/Path	Owner	Format	Retention
GAI Profiles (per use case)	GOV/MAP/ MES/MAN	AI/Profiles/	RAI Lead	PDF/MD	Current + 1 yr
Policy, Roles, Gates & Minutes	GOVERN	AI/Governance/	RAI Lead	PDF	3 yrs
AIBOM/Registry Entries	GOVERN	AI/Registry/	Model Owner	CSV/MD	Current + 1 yr
Data Inventory & Lineage	MAP	AI/Data/	Data Owner	CSV/PDF	3 yrs
Data Rights & Provenance Attestations	MAP	AI/Data/Rights/	Legal/Data	PDF	3 yrs
RAG Source Allow-List & Rules	MAP	AI/RAG/	Model Owner	MD/PDF	Current + 1 yr
Harm Analysis & Affected Parties	MAP	AI/Risk/	Model Owner	PDF	3 yrs
TEVV Plans & Metrics Catalog	MEASURE	AI/TEVV/	TEVV Lead	PDF/MD	3 yrs
Safety/Hallucination/Privacy/ IP Eval Reports	MEASURE	AI/Evals/	TEVV Lead	PDF/CSV	3 yrs
Red-Team Reports (GAI)	MEASURE/ MANAGE	AI/RedTeam/	Sec/TEVV	PDF	3 yrs
Guardrail & Tool-Use Configs/ Approvals	MANAGE	AI/Controls/	Model Owner	JSON/PDF	2 yrs
Monitoring Dashboards & Alerts	MANAGE	AI/Mon/	SRE/Model Ops	PNG/CSV	1–2 yrs
Incident Logs & AARs (GAI taxonomy)	MANAGE	IR/AI/	IR Lead	PDF	3 yrs
Pre-Launch Gate Checklist & Minutes	GOV/ MANAGE	AI/Governance/ Gates/	RAI Lead	PDF	3 yrs
Rollback Test Evidence	MANAGE	Change/AI/Rollback/	Ops	PDF/CSV	2 yrs
Supplier Responsibility Matrices	GOV/ MANAGE	TPRM/AI/	Procurement	PDF/XLSX	Active + 1 yr
Change Tickets & Approvals	MANAGE	Change/AI/	Ops	PDF/CSV	2 yrs

9. Continuous Monitoring

- **Daily:** Ingestion/log health; unsafe content flags; jailbreak/injection alerts; tool-abuse anomalies; PII/IP flags; drift signals; failed eval jobs.
- **Weekly:** Bias slice deltas; RAG source changes; secret/certificate checks; prompt/policy change review.
- **Monthly:** Threshold review; retraining or safety-rule updates; dependency updates; provider notices.
- **Quarterly:** Full evaluation reruns (High tier); red-team exercise; management KPI review.
- **Automation:** Auto-ticket on threshold breach (ASR, hallucination, leakage, IP duplication), missing logs, or unreviewed alerts; track to closure.

10. Remediation & Risk Acceptance Workflow

- Identify → Log (owner, severity, due date, milestones) → Treat (guardrail/prompt/tool changes; retraining; data/RAG fixes; UI/UX changes; vendor updates) → Verify with evidence → Report to governance.
- Risk acceptance only by the Executive GAI Risk Owner with explicit expiry, conditions, and follow-up measurements; keep residual risks visible in the GAI Profile and governance reviews.

11. External Providers & Model/Service Dependencies (GAI-Specific)

Inventory foundation models, hosted APIs, safety toolkits, plug-ins/function hubs, RAG providers, vector DBs, content moderation, watermark/provenance services. For each: shared responsibilities, evaluation rights, update transparency, telemetry access, incident notice, IP indemnity stance, data usage/retention, provenance/watermark verification, and subcontractor flow-downs. Maintain on-file evaluations of provider safety/robustness claims.

12. Training & Awareness (Role-Based)

Tracks for workforce, business owners, DS/ML & prompt engineers, safety/red-teamers, SRE/Model Ops, Legal/IP & Ethics, Procurement, leadership. Onboarding before access; annual refresh; targeted refreshers post-incident. Metrics: completion ≥ 98% and declining repeat findings.

13. Change Management for AI Lifecycle Impact

Gate changes to datasets, **system prompts/policies**, guardrails, tool lists/functions, RAG sources, and dependencies. Re-run material TEVV; update GAI Profile and registry; obtain gate approval before production; maintain rollback artifacts and signed versioned configs.

14. Evidence Sampling Plans (Internal QA)

- **Governance:** Sample three launches for gate compliance and AIBOM completeness.
- **MAP:** Inspect two profiles for rights/provenance and RAG governance.
- **MEASURE:** Reproduce one evaluation end-to-end; review three bias slices; re-run jailbreak battery; run a privacy/IP duplication check.
- **MANAGE:** Simulate one rollback and one GAI incident per quarter for High-tier use cases; verify detection and response SLAs.
- **Third-Party:** Review two supplier matrices and last safety/update notice.

15. Common Pitfalls

Launching with only “helpfulness” checks; missing rights for training/fine-tuning; unmanaged RAG sources; no jailbreak/PII/IP tests; tool/function abuse paths; logging prompts without PII safeguards; no provenance stance; inability to rollback.

16. Quick Reference Summary

Area	Core Artifacts	Examples
GOVERN	Policy; roles; gates; AIBOM	Gate minutes; signed prompt/policy versions
MAP	Profiles; rights; RAG rules; harms	Consent/licenses; allow-list; mitigation owners
MEASURE	Safety/Factuality/Privacy/IP suite	ASR; hallucination; leakage; duplication
MANAGE	Guardrails; monitoring; IR	Filters; tool lists; provenance; rollback
Third-Party	Responsibility matrices	Eval rights; incident notice; transparency

17. Self-Assessment & Leadership Attestation

Use status: **Compliant (C) / Partially Compliant (PC) / Not Compliant (NC) / Not Applicable (N/A)**.

Tracker (illustrative):

Function/Outcome	Status	Rationale (for N/A or gaps)	Evidence Link	Owner	Action ID
MEASURE – Safety & Hallucination	PC	ASR 2.4% > 1% High-tier	AI/Evals/2025Q3.pdf	TEVV Lead	ACT-2025-071
MANAGE – Guardrails & Tool-Use	C	Allow/deny lists enforced; drill pass	AI/Controls/ToolUse.md	Model Owner	—
MAP – Rights & RAG Governance	C	Licenses verified; sources approved	AI/RAG/AllowList.md	Legal/Data	—

Leadership checks (explicit):

- □ AIBOM/Registry entry complete
- □ Independent evaluation completed (per tier)
- □ Pre-launch gate passed; residual risks accepted by Executive GAI Risk Owner
- □ Monitoring + rollback tested and operational

18. References & Resources

NIST AI RMF 1.0 (NIST AI 100-1)
https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

NIST AI RMF Playbook
https://airc.nist.gov/airmf-resources/playbook/

ISO/IEC 42001:2023 (AI Management System)
https://www.iso.org/standard/81230.html

ISO/IEC 23894:2023 (AI Risk Management)
https://www.iso.org/standard/77304.html

NIST SP 800-218 (Secure Software Development Framework)
https://csrc.nist.gov/publications/detail/sp/800-218/final

NIST Privacy Framework
https://www.nist.gov/privacy-framework

OWASP Top 10 for LLM Applications
https://owasp.org/www-project-top-10-for-large-language-model-applications/

ENISA AI Threat Landscape
https://www.enisa.europa.eu/publications

Content Authenticity Initiative / C2PA (Provenance)
https://c2pa.org

Apptega Product Features

16+ Security Frameworks

One-Click Reporting

Automated Alerts & Notifications

API & Application Connectors

Automated Framework Crosswalking

Real-Time Compliance Scoring

Restricted Auditor View

Single Sign-On Connectivity

Policy & Plan Templates

Automated Risk Assessments

Document Repository for Artifacts

Multi-Tenant Environment

Apptega

About Apptega

A perennial G2 leader across various cybersecurity categories, Apptega is the cybersecurity and compliance platform purpose-built for security providers who are growing lucrative security practices, creating stickier customer relationships, and winning more business from competitors.

With Apptega's end-to-end platform, IT providers can:

1. Build continuous compliance offerings that set them apart, retain customers, and grow their ARR.
2. Build and manage world-class cybersecurity and compliance programs for their clients.
3. Increase the capacity and efficiency of their existing team so they can service 2–3× more customers.
4. Assess, measure, and manage their own security programs against a standard set of industry-leading frameworks.

To learn more, visit apptega.com

Visit apptega.com