THE 2026 ENTERPRISE AI SECURITY CHECKLIST FOR CISOs

AI Infrastructure Security Hub

elvis.hk · Q3 2026 · 55 Controls · 11 Domains

Generated: June 2026

55	11	6
Total Controls	Security Domains	P1-Critical Domains

Table of Contents

DOMAIN 01 — LLM PROCUREMENT & VENDOR ASSESSMENT

P1-Critical | Applies to: **All LLMs** | 5 controls

- Verify vendor's responsible disclosure and security advisory publication process
- Review model card for known limitations, bias evaluations, and safety benchmarks
- Assess data residency, training data provenance, and opt-out controls
- Confirm SOC 2 Type II / ISO 27001 certification for the API provider
- Evaluate fine-tuning data supply chain risks — third-party dataset vetting

DOMAIN 02 — SYSTEM PROMPT & CONTEXT SECURITY

P1-Critical | Applies to: **All LLMs** | 5 controls

- Classify system prompt as confidential — never expose verbatim to users or logs
- Implement prompt injection detection layer with input validation and normalization
- Define strict output schemas to prevent data exfiltration via generation
- Restrict tool/function call permissions to least-privilege principles
- Log and alert on anomalous instruction patterns (e.g. repeated jailbreak probes)

DOMAIN 03 — ACCESS CONTROL & IDENTITY

P1-Critical | Applies to: **Both** | 5 controls

- Apply per-user API key scoping — prohibit shared org-wide API tokens
- Enforce MFA on all AI platform admin accounts and developer consoles
- Rotate API keys on a 90-day schedule; revoke immediately on staff offboarding
- Implement RBAC for AI feature access (model tier, context window size, tool use)
- Audit API key usage logs weekly and set anomaly alerts on unusual volumes

DOMAIN 04 — RAG & RETRIEVAL PIPELINE SECURITY

P1-Critical | Applies to: **Both** | 5 controls

- Sanitize and validate all documents before ingestion into the vector store
- Implement document-level access control in retrieval layer (per-tenant isolation)
- Monitor for RAG poisoning — alert on unexpected document mutations or additions
- Isolate retrieval namespace per tenant in multi-tenant deployments
- Validate and score retrieved context relevance before injecting into prompt

DOMAIN 05 — OUTPUT VALIDATION & CONTENT FILTERING

P1-Critical | Applies to: **Both** | 5 controls

- Implement output classifiers for PII, secrets, and harmful content detection
- Block markdown/code injection in rendered outputs — sanitize HTML and JS
- Validate structured outputs against strict JSON schema before downstream use
- Apply rate limiting on generation endpoints to prevent abuse and data harvesting
- Log all model outputs for audit trail and post-incident forensics

DOMAIN 06 — AGENTIC & AUTONOMOUS WORKFLOW SECURITY

P1-Critical | Applies to: **Both** | 5 controls

- Enforce human-in-the-loop approval for all irreversible agentic actions
- Sandbox code execution environments — no access to host filesystem or network
- Limit agent memory scope — enforce no cross-session or cross-user context leakage
- Define and enforce agent capability boundaries explicitly in system prompt
- Monitor agent action logs for privilege escalation and anomalous tool invocations

DOMAIN 07 — MONITORING, LOGGING & INCIDENT RESPONSE

P1-Critical | Applies to: **Both** | 5 controls

- Enable full prompt/completion audit logging with PII masking and retention policy
- Set anomaly detection on tokens-per-request, request rate, and refusal rate
- Define and test AI-specific incident response runbook with escalation paths
- Conduct quarterly red-team exercises against production AI endpoints
- Subscribe to model provider security bulletins (Anthropic, OpenAI changelog)

DOMAIN 08 — SUPPLY CHAIN & THIRD-PARTY MODEL RISK

P2-High | Applies to: **Both** | 5 controls

- Pin model versions — do not auto-upgrade without security review and regression testing
- Vet all LangChain, LlamaIndex, and custom plugins before production deployment
- Review open-source model weights for known backdoors before fine-tuning
- Assess third-party AI vendor sub-processors under your vendor risk framework
- Maintain SBOM (Software Bill of Materials) for all AI/ML dependencies

DOMAIN 09 — DATA PRIVACY & COMPLIANCE

P2-High | Applies to: **Both** | 5 controls

- Confirm API provider does not train on customer data without explicit consent
- Implement data minimisation — strip PII before sending to third-party LLM APIs
- Map AI data flows for GDPR/PDPO Article 30 records of processing activities
- Assess cross-border data transfer legality for cloud LLM API calls
- Classify AI-generated outputs under your organization's data classification policy

DOMAIN 10 — CLAUDE-SPECIFIC CONTROLS (ANTHROPIC)

P2-High | Applies to: **Claude** | 5 controls

- Use Constitutional AI system prompt patterns from Anthropic's official documentation
- Implement Anthropic's recommended prompt injection hardening via structured system prompt
- Review Claude's tool use security guidance before enabling function calls in production
- Test quarterly against known Claude-specific jailbreak patterns (JB-2026-Q3-001 to 008)
- Monitor Anthropic's responsible scaling policy updates and model safety reports

DOMAIN 11 — OPENAI-SPECIFIC CONTROLS

P2-High | Applies to: OpenAI | 5 controls

- [ ] Enable OpenAI moderation API on all user-facing text generation endpoints
- [ ] Use Structured Outputs (JSON mode) to prevent instruction injection via generation
- [ ] Review OpenAI's usage policies and configure custom safety instructions per deployment
- [ ] Test GPT-4o vision inputs for prompt injection via adversarial image content
- [ ] Monitor OpenAI's deprecated model timelines and security changelog for CVE alerts

AI Infrastructure Security Hub

elvis.hk

This checklist covers 55 security controls across 11 domains for enterprise AI deployments of Anthropic Claude and OpenAI GPT models.

Sources: OWASP Top 10 for LLMs · Anthropic Research · OpenAI Safety · elvis.hk CVE Feeds

Q3 2026 · Not legal advice · elvis.hk