Shadow AI Operational Maturity Model (SAOMM)

1. Scope of This Document

This condensed model provides a unified, enterprise-ready view of AI governance, security, and operational maturity. It consolidates the most authoritative frameworks in the industry into a single, structured assessment model. The document covers:

AI Governance & Responsible AI
AI Strategy, Organization & Resourcing
Data Security & Model Security
LLM Security & Agentic AI Security
Monitoring, Detection & Observability
MLOps, Deployment & Infrastructure Security
MITRE ATLAS Tactics (Condensed)
MITRE ATLAS Mitigations (Condensed)

Each control includes five maturity levels (Level 5 → Level 1), with descriptions and examples to support consistent scoring across teams.

Shadow AI Operational Domains

Domain	Operational Focus
AI Visibility	Discovery of AI SaaS usage and workflow spread
AI Governance	Sanctioned vs unsanctioned AI usage
AI Telemetry	Visibility into prompts, uploads, and AI interactions
AI Identity	OAuth grants, delegated access, AI-linked identities
AI Trust Boundaries	Browser extensions, mediation layers, workflow drift
AI Detection Engineering	Behavioral analytics and AI abuse detection
AI Data Protection	Prompt DLP and sensitive data exposure
AI Incident Response	Investigation and containment of AI-mediated events

Level 1: No Visibility

AI Traffic Ratio unknown
Coverage < 20%
No CASB tagging for AI domains
MTTD undefined
👉 KPI signature:

AI visibility = 0–20%

Level 2: Domain-Level Detection
AI domains identified (ChatGPT, Claude, Copilot)
AI Traffic Ratio measurable
CASB blocking basic tools
👉 KPI signature:
Coverage 20–50%
MTTD > 1 hour

Level 3: Prompt-Level Detection
Prompt DLP enabled
Sensitive prompt detection active
Shadow AI incidents correlated to users
👉 KPI signature:
Prompt DLP hit rate measurable
Coverage 50–75%
MTTD < 30 min

Level 4: AI Governance Automation
Policy enforcement automated
Risk scoring per AI session
Auto-blocking unsafe prompts/tools
👉 KPI signature:
MTTR < 15 min
Policy exception growth stabilized
Coverage 75–90%

Level 5: AI-Native Zero Trust
Every AI request authenticated + contextualized
Prompt-level authorization (not just domain blocking)
Continuous risk scoring per interaction
👉 KPI signature:
Coverage > 90%
Near real-time MTTD (< 5 min)
AI risk score per session enforced

2. Source Frameworks & References

This model synthesizes controls from the most widely recognized AI security and governance frameworks. Each control is mapped to one or more of the following sources:

Primary Sources

MITRE ATLAS — Adversarial Threat Landscape for Artificial-Intelligence Systems (Tactics, Techniques, Mitigations)
MITRE AI Maturity Model — Organizational AI capability and governance maturity
OWASP Top 10 for LLM Applications — LLM-specific security risks
OWASP Top 10 for Agentic AI Systems — Agent autonomy, tool-use, and memory risks

Secondary Sources

NIST AI Risk Management Framework (AI RMF)
ISO/IEC 42001 (AI Management System Standard)
Industry best practices from cloud providers (Azure, AWS, GCP)

This document does not reproduce copyrighted material; instead, it provides derived controls aligned to these frameworks.

3. How to Use This Model

This maturity model is designed to be used by:

CISOs & Security Architects
AI Governance & Responsible AI teams
MLOps & Platform Engineering
Red Teams & AI Security Testing groups
Compliance, Audit, and Risk Management

Recommended Usage Workflow

Identify relevant control categories for your AI systems.

For each control, score your current maturity level (Level 1–5).
Use the examples to ensure consistent interpretation across teams.
Document gaps and define target maturity levels.
Build a roadmap to close gaps across governance, security, and operations.
Reassess quarterly or after major AI system changes.

Scoring Guidance

Level 5 = Automated, optimized, continuously monitored
Level 4 = Managed, measurable, consistently enforced
Level 3 = Defined, documented, repeatable
Level 2 = Developing, inconsistent, partially implemented
Level 1 = Initial, ad-hoc, no formal controls

Intended Outcomes

Establish a baseline for AI governance and security maturity
Identify high-risk gaps in LLM, agent, and model security
Provide a roadmap for enterprise AI hardening
Enable consistent scoring across business units
Support compliance with emerging AI regulations

4. Document Structure

The remainder of this document is organized into batches of controls, each with:
Control Category
Control Name
Maturity Level (5 → 1)
Description

Example
Score (to be filled in)
Comments (to be filled in)
This structure ensures the model is Excel-ready, auditable, and easy to operationalize.

5. Related Sections You May Want to Explore Next

Governance Controls
LLM Security Controls
Agentic AI Controls
ATLAS Tactics
ATLAS Mitigations

Author: Julie BRUNIAS

Role: Assembler and curator of the Condensed AI Security & Maturity Model

Note:

This document synthesizes and reorganizes controls derived from publicly available frameworks including MITRE ATLAS, MITRE AI Maturity Model, OWASP LLM Top 10, OWASP Agentic AI Top 10, NIST AI RMF, and ISO/IEC 42001. The structure, maturity levels, examples, and unified control taxonomy were created by the author.

Items	Average score
Governance	0
Strategy	0
Organization	0
Detection	0
Monitoring	0
Incident Response	0
ML-Ops	0
Deployment	0
Infrastructure	0
Resources	0
Data Security	0
Model Security	0
LLM Security	0
Agent Security	0
LLM Agent Guardrails	0
ATLAS Tactics Coverage	0
ATLAS Mitigation Coverage	0
Total Average Score	0

Governance

Control Category	Control Name	Maturity Level	Description / Criteria	Example	Score	Comments
Governance	AI Program Governance	Level 5 – Optimized	Governance is automated, continuously monitored, and integrated with enterprise risk systems.	AI governance platform auto-flags non-compliant models before deployment.		
Governance	AI Program Governance	Level 4 – Managed	Governance processes are standardized, measured, and enforced across all teams.	Mandatory AI review gates in CI/CD.		
Governance	AI Program Governance	Level 3 – Defined	Formal governance roles, policies, and processes exist and are documented.	AI steering committee, RACI, policies.		
Governance	AI Program Governance	Level 2 – Developing	Some governance exists but is inconsistent or team-specific.	One team uses a model registry, others don't.		
Governance	AI Program Governance	Level 1 – Initial	No formal governance; ad-hoc decisions.	No inventory of AI systems.		
			Total AI Program Governance		#DIV/0!	
Governance	Responsible AI	Level 5 – Optimized	Automated fairness, bias, and transparency checks run continuously.	Bias detection triggers auto-rollback.		
Governance	Responsible AI	Level 4 – Managed	Responsible AI KPIs are monitored and audited.	Quarterly fairness audits.		
Governance	Responsible AI	Level 3 – Defined	Responsible AI standards and documentation are formalized.	Model cards, transparency reports.		
Governance	Responsible AI	Level 2 – Developing	Basic guidelines exist but are not enforced.	Ethical checklist used by some teams.		
Governance	Responsible AI	Level 1 – Initial	No ethical or fairness considerations.	No bias testing.		
			Total Responsible AI		#DIV/0!	
Governance	AI Risk Management	Level 5 – Optimized	Automated risk scoring and mitigation workflows.	Real-time risk scoring dashboard.		

Governance	AI Risk Management	Level 4 – Managed	Continuous monitoring of AI risks across lifecycle.	Drift alerts tied to risk register.	
Governance	AI Risk Management	Level 3 – Defined	Formal AI risk register and assessment process.	Risk scoring templates.	
Governance	AI Risk Management	Level 2 – Developing	Ad-hoc risk reviews.	One-off risk assessments.	
Governance	AI Risk Management	Level 1 – Initial	No AI-specific risk management.	No tracking of AI risks.	
			Total AI Risk Management		#DIV/0!
Governance	AI Policy & Compliance	Level 5 – Optimized	Policies enforced automatically through tooling.	Automated policy enforcement in pipelines.	
Governance	AI Policy & Compliance	Level 4 – Managed	Compliance is monitored and audited regularly.	Quarterly compliance audits.	
Governance	AI Policy & Compliance	Level 3 – Defined	AI policies exist and are communicated.	AI acceptable-use policy.	
Governance	AI Policy & Compliance	Level 2 – Developing	Draft policies exist but are not enforced.	Draft RAI guidelines.	
Governance	AI Policy & Compliance	Level 1 – Initial	No AI policies.	No rules for model usage.	
			Total AI Policy/Compliance		#DIV/0!
Governance	AI Inventory & Registry	Level 5 – Optimized	Automated discovery of all AI systems; real-time updates.	Auto-scanning cloud for unregistered models.	
Governance	AI Inventory & Registry	Level 4 – Managed	Registry is mandatory and enforced.	CI/CD blocks unregistered models.	
Governance	AI Inventory & Registry	Level 3 – Defined	Centralized model registry exists.	MLflow, SageMaker, Azure registry.	
Governance	AI Inventory & Registry	Level 2 – Developing	Partial inventory exists.	Spreadsheet of models.	
Governance	AI Inventory & Registry	Level 1 – Initial	No inventory.	No idea what models exist.	
			Total AI Inventory/Registry		#DIV/0!

Strategy

Control Category	Control Name	Maturity Level	Description / Criteria	Example	Score	Comments
Strategy	AI Strategy & Vision	Level 5 – Optimized	AI strategy is continuously refined using real-time performance, risk, and value metrics.	Strategy auto-adjusts based on model ROI dashboards.		
Strategy	AI Strategy & Vision	Level 4 – Managed	AI strategy is measurable, aligned to business goals, and reviewed regularly.	Annual AI strategy review with KPIs.		
Strategy	AI Strategy & Vision	Level 3 – Defined	Documented AI strategy exists and is communicated across the organization.	AI strategy published internally.		
Strategy	AI Strategy & Vision	Level 2 – Developing	AI strategy is emerging but incomplete or siloed.	One business unit has an AI roadmap.		
Strategy	AI Strategy & Vision	Level 1 – Initial	No AI strategy; ad-hoc experimentation.	Teams build models without direction.		
			Total AI Strategy/Vision		#DIV/0!	
Strategy	AI Value Realization	Level 5 – Optimized	Automated measurement of AI value, cost, and impact across the portfolio.	Real-time ROI dashboards for all models.		
Strategy	AI Value Realization	Level 4 – Managed	AI value is measured consistently with standardized KPIs.	Cost savings tracked quarterly.		
Strategy	AI Value Realization	Level 3 – Defined	Value measurement frameworks exist.	Business case templates for AI projects.		
Strategy	AI Value Realization	Level 2 – Developing	Some value tracking exists but is inconsistent.	One team tracks model performance.		
Strategy	AI Value Realization	Level 1 – Initial	No measurement of AI value.	No tracking of AI impact.		
			Total AI Value Realization		#DIV/0!	

Strategy	AI Portfolio Management	Level 5 – Optimized	Portfolio decisions are automated using risk, performance, and value signals.	Auto-retirement of low-value models.	
Strategy	AI Portfolio Management	Level 4 – Managed	Portfolio is actively managed with clear prioritization.	Quarterly portfolio reviews.	
Strategy	AI Portfolio Management	Level 3 – Defined	AI portfolio is documented and tracked.	Centralized list of all AI initiatives.	
Strategy	AI Portfolio Management	Level 2 – Developing	Partial portfolio visibility.	Some projects tracked, others not.	
Strategy	AI Portfolio Management	Level 1 – Initial	No portfolio management.	No visibility into AI initiatives.	
			Total AI Portfolio		#DIV/0!

Organization

Control Category	Control Name	Maturity Level	Description / Criteria	Example	Score	Comments
Organization	AI Organizational Structure	Level 5 – Optimized	AI org structure adapts dynamically to business needs; federated + centralized hybrid.	AI Center of Excellence + embedded teams.		
Organization	AI Organizational Structure	Level 4 – Managed	Clear roles, responsibilities, and cross-functional alignment.	Central AI team with defined charters.		
Organization	AI Organizational Structure	Level 3 – Defined	Formal AI org structure exists.	AI lead, MLOps team, governance roles.		
Organization	AI Organizational Structure	Level 2 – Developing	Some roles exist but unclear or siloed.	One data scientist per team.		
Organization	AI Organizational Structure	Level 1 – Initial	No AI roles or structure.	No dedicated AI staff.		
			Total AI Organizational		#DIV/0!	
Organization	AI Skills & Workforce	Level 5 – Optimized	Workforce skills are continuously assessed and training is automated/personalized.	AI skill gap analysis auto-generated.		
Organization	AI Skills & Workforce	Level 4 – Managed	Enterprise-wide AI training programs exist and are measured.	Annual AI upskilling program.		
Organization	AI Skills & Workforce	Level 3 – Defined	AI training is available and structured.	AI bootcamps, internal courses.		
Organization	AI Skills & Workforce	Level 2 – Developing	Some training exists but is inconsistent.	One team uses Coursera.		
Organization	AI Skills & Workforce	Level 1 – Initial	No AI training.	No AI education programs.		
			Total AI Skills/Workforce		#DIV/0!	

Detection

Control Category	Control Name	Maturity Level	Description / Criteria	Example	Score	Comments
Detection	Domain-Level Detection	Level 5 – Optimized	Autonomous detection of AI threats across domains with risk scoring.	Auto-flagging anomalous inference patterns.		
Detection	Domain-Level Detection	Level 4 – Managed	ML-based anomaly detection across AI systems.	Behavioral analytics for model misuse.		
Detection	Domain-Level Detection	Level 3 – Defined	Standardized detection rules and thresholds.	Thresholds for unusual API usage.		
Detection	Domain-Level Detection	Level 2 – Developing	Basic anomaly detection.	Simple volume-based alerts.		
Detection	Domain-Level Detection	Level 1 – Initial	No detection.	No alerts for abnormal behavior.		
			Total Domain Lvl Detection		#DIV/0!	
Detection	Prompt-Level Detection	Level 5 – Optimized	Autonomous detection of malicious prompts using AI-native Zero Trust.	Auto-blocking indirect prompt injection.		
Detection	Prompt-Level Detection	Level 4 – Managed	ML-based detection of harmful prompt patterns.	Prompt anomaly classifier.		
Detection	Prompt-Level Detection	Level 3 – Defined	Standardized prompt validation rules.	Prompt allow/deny lists.		
Detection	Prompt-Level Detection	Level 2 – Developing	Basic keyword-based detection.	Regex for jailbreak attempts.		
Detection	Prompt-Level Detection	Level 1 – Initial	No prompt detection.	Any prompt accepted.		
			Total Prompt Lvl Detection		#DIV/0!	

Detection	AI Threat Analytics	Level 5 – Optimized	Autonomous threat correlation across all AI systems.	AI SOC auto-correlates LLM attacks.		
Detection	AI Threat Analytics	Level 4 – Managed	Continuous threat analytics with dashboards.	LLM threat analytics platform.		
Detection	AI Threat Analytics	Level 3 – Defined	Documented threat analytics processes.	ATLAS-aligned threat mapping.		
Detection	AI Threat Analytics	Level 2 – Developing	Partial threat analysis.	Manual review of logs.		
Detection	AI Threat Analytics	Level 1 – Initial	No threat analytics.	No visibility into AI threats.		
				Total AI Threat Analytics	#DIV/0!	

Monitoring

Control Category	Control Name	Maturity Level	Description / Criteria	Example	Score	Comments
Monitoring	AI System Observability	Level 5 – Optimized	Full-stack AI observability with autonomous anomaly detection and automated remediation.	System auto-rolls back a failing model.		
Monitoring	AI System Observability	Level 4 – Managed	Continuous monitoring of inputs, outputs, latency, and performance.	Real-time dashboards for all models.		
Monitoring	AI System Observability	Level 3 – Defined	Standardized observability metrics and logging.	Model latency, accuracy, drift metrics.		
Monitoring	AI System Observability	Level 2 – Developing	Partial monitoring; inconsistent metrics.	One team tracks inference latency.		
Monitoring	AI System Observability	Level 1 – Initial	No monitoring.	No visibility into model behavior.		
				Total AI System Obs	#DIV/0!	
Monitoring	Agent Behavior Monitoring	Level 5 – Optimized	Autonomous detection of risky agent behavior with automated containment.	Agent auto-sandboxed during anomalies.		
Monitoring	Agent Behavior Monitoring	Level 4 – Managed	Continuous monitoring of agent actions and tool calls.	Alerts for unusual tool sequences.		
Monitoring	Agent Behavior Monitoring	Level 3 – Defined	Standardized behavior logging.	Logs of all agent actions.		
Monitoring	Agent Behavior Monitoring	Level 2 – Developing	Partial monitoring.	Logs only some agent actions.		
Monitoring	Agent Behavior Monitoring	Level 1 – Initial	No monitoring.	Agent actions invisible.		
				Total Agent Behavior Monitoring	#DIV/0!	

Monitoring	Model Drift Detection	Level 5 – Optimized	Autonomous drift detection with automated retraining or rollback.	Auto-rollback when drift exceeds threshold.	
Monitoring	Model Drift Detection	Level 4 – Managed	Continuous drift monitoring with alerts.	Drift dashboard with thresholds.	
Monitoring	Model Drift Detection	Level 3 – Defined	Documented drift detection processes.	Statistical drift tests.	
Monitoring	Model Drift Detection	Level 2 – Developing	Partial drift checks.	Occasional manual drift analysis.	
Monitoring	Model Drift Detection	Level 1 – Initial	No drift detection.	Model performance degrades unnoticed.	
			Total Model Drift Detection		#DIV/0!
Monitoring	Audit Logging & Traceability	Level 5 – Optimized	Immutable, automated audit logs with full traceability and anomaly detection.	Blockchain-backed audit logs.	
Monitoring	Audit Logging & Traceability	Level 4 – Managed	Continuous audit log monitoring.	Alerts for suspicious log entries.	
Monitoring	Audit Logging & Traceability	Level 3 – Defined	Standardized logging across all AI systems.	Input/output logging.	
Monitoring	Audit Logging & Traceability	Level 2 – Developing	Partial logging.	Logs only some model interactions.	
Monitoring	Audit Logging & Traceability	Level 1 – Initial	No logging.	No record of model activity.	
			Total Audit Logging Traceability		#DIV/0!

Incident Response

Control Category	Control Name	Maturity Level	Description / Criteria	Example	Score	Comments
Incident Response	AI Incident Response	Level 5 – Optimized	Autonomous containment and recovery workflows for AI incidents.	Auto-quarantine of compromised model.		
Incident Response	AI Incident Response	Level 4 – Managed	Continuous monitoring and rehearsed AI-specific IR plans.	AI red-team exercises.		
Incident Response	AI Incident Response	Level 3 – Defined	Documented AI incident response playbooks.	ATLAS-aligned IR procedures.		
Incident Response	AI Incident Response	Level 2 – Developing	Partial IR processes.	Manual response to AI failures.		
Incident Response	AI Incident Response	Level 1 – Initial	No AI-specific incident response.	No plan for LLM compromise.		
				Total AI Incident Response	#DIV/0!	

MLOps

Control Category	Control Name	Maturity Level	Description / Criteria	Example	Score	Comments
MLOps	Model Lifecycle Management	Level 5 – Optimized	Autonomous lifecycle orchestration with automated retraining, rollback, and retirement.	Auto-retirement of stale models.		
MLOps	Model Lifecycle Management	Level 4 – Managed	Continuous lifecycle monitoring with standardized workflows.	CI/CD triggers retraining pipelines.		
MLOps	Model Lifecycle Management	Level 3 – Defined	Documented lifecycle processes for development → deployment → monitoring → retirement.	Versioning, approval gates.		
MLOps	Model Lifecycle Management	Level 2 – Developing	Partial lifecycle processes.	Some models versioned, others not.		
MLOps	Model Lifecycle Management	Level 1 – Initial	No lifecycle management.	Models deployed manually.		
			Total Model Lifecycle		#DIV/0!	
MLOps	CI/CD for AI	Level 5 – Optimized	Fully automated CI/CD with AI-specific validation, security checks, and governance gates.	Auto-block deployment if safety tests fail.		
MLOps	CI/CD for AI	Level 4 – Managed	Continuous integration and deployment with automated tests.	Automated model validation suite.		
MLOps	CI/CD for AI	Level 3 – Defined	Standardized CI/CD pipelines for AI.	Model registry + deployment pipeline.		
MLOps	CI/CD for AI	Level 2 – Developing	Partial automation.	Manual deployment with some scripts.		
MLOps	CI/CD for AI	Level 1 – Initial	No CI/CD.	Manual model deployment.		
			Total CI/CD for AI		#DIV/0!	

Deployment

Control Category	Control Name	Maturity Level	Description / Criteria	Example	Score	Comments
Deployment	Secure Model Deployment	Level 5 – Optimized	Autonomous deployment validation with Zero Trust enforcement.	Auto-deny deployment of non-compliant models.		
Deployment	Secure Model Deployment	Level 4 – Managed	Deployment security checks enforced consistently.	Security scanning in deployment pipeline.		
Deployment	Secure Model Deployment	Level 3 – Defined	Documented deployment security standards.	Deployment checklist.		
Deployment	Secure Model Deployment	Level 2 – Developing	Partial deployment controls.	Some models deployed securely.		
Deployment	Secure Model Deployment	Level 1 – Initial	No deployment security.	Models deployed without review.		
				Total Secure Model Deployment	#DIV/0!	

Infrastructure

Control Category	Control Name	Maturity Level	Description / Criteria	Example	Score	Comments
Infrastructure	Environment Hardening	Level 5 – Optimized	Autonomous environment hardening with continuous compliance scanning.	Auto-patching of ML environments.		
Infrastructure	Environment Hardening	Level 4 – Managed	Continuous monitoring of environment security.	Vulnerability scanning for ML runtimes.		
Infrastructure	Environment Hardening	Level 3 – Defined	Standardized environment hardening procedures.	Hardened containers for inference.		
Infrastructure	Environment Hardening	Level 2 – Developing	Partial hardening.	Some secure configs applied.		
Infrastructure	Environment Hardening	Level 1 – Initial	No hardening.	Default insecure environments.		
			Total Environment Hardening		#DIV/0!	
Infrastructure	Sandboxing & Isolation	Level 5 – Optimized	Zero-Trust execution isolation with autonomous sandboxing.	Auto-sandboxing during suspicious behavior.		
Infrastructure	Sandboxing & Isolation	Level 4 – Managed	Continuous monitoring of sandbox integrity.	Alerts for sandbox escapes.		
Infrastructure	Sandboxing & Isolation	Level 3 – Defined	Standardized sandboxing for all models/agents.	Containerized inference.		
Infrastructure	Sandboxing & Isolation	Level 2 – Developing	Partial sandboxing.	Some models run in isolated VMs.		
Infrastructure	Sandboxing & Isolation	Level 1 – Initial	No sandboxing.	Models run with full system access.		
			Total Sandboxing/Isolation		#DIV/0!	
Infrastructure	Kill Switch	Level 5 – Optimized	Autonomous kill switch triggered by risk scoring or anomaly detection.	Auto-shutdown of rogue agent.		
Infrastructure	Kill Switch	Level 4 – Managed	Automated kill switch with monitoring.	Kill switch tied to safety metrics.		

Infrastructure	Kill Switch	Level 3 – Defined	Documented kill switch procedures.	Manual kill switch in UI.	
Infrastructure	Kill Switch	Level 2 – Developing	Partial kill switch capability.	CLI-based shutdown.	
Infrastructure	Kill Switch	Level 1 – Initial	No kill switch.	No way to stop a runaway model.	
			Total Kill Switch		#DIV/0!
Infrastructure	AI Zero Trust	Level 5 – Optimized	Full AI-native Zero Trust with continuous authentication, authorization, and risk scoring.	Dynamic access revocation for models.	
Infrastructure	AI Zero Trust	Level 4 – Managed	Zero Trust enforced across AI systems.	Identity-aware inference endpoints.	
Infrastructure	AI Zero Trust	Level 3 – Defined	Zero Trust policies documented.	Access segmentation for AI workloads.	
Infrastructure	AI Zero Trust	Level 2 – Developing	Partial Zero Trust controls.	Some identity checks.	
Infrastructure	AI Zero Trust	Level 1 – Initial	No Zero Trust.	Flat network, no segmentation.	
			Total AI Zero Trust		#DIV/0!
Infrastructure	Dependency & Environment Security	Level 5 – Optimized	Autonomous dependency verification and continuous SBOM monitoring.	Auto-flagging compromised ML libraries.	
Infrastructure	Dependency & Environment Security	Level 4 – Managed	Continuous scanning of ML dependencies.	Automated CVE scanning.	
Infrastructure	Dependency & Environment Security	Level 3 – Defined	SBOM required for all models.	MLflow + SBOM integration.	
Infrastructure	Dependency & Environment Security	Level 2 – Developing	Partial dependency tracking.	Some libraries tracked manually.	
Infrastructure	Dependency & Environment Security	Level 1 – Initial	No supply chain controls.	Unknown ML dependencies.	
			Total Dependency/Environment Security		#DIV/0!

Resources

Control Category	Control Name	Maturity Level	Description / Criteria	Example	Score	Comments
Resources	AI Budgeting & Funding	Level 5 – Optimized	Funding is dynamic, value-based, and automated.	Funding shifts based on model ROI.		
Resources	AI Budgeting & Funding	Level 4 – Managed	AI budget is planned, tracked, and aligned to strategy.	Annual AI budget cycle.		
Resources	AI Budgeting & Funding	Level 3 – Defined	AI funding model exists.	Dedicated AI budget line.		
Resources	AI Budgeting & Funding	Level 2 – Developing	Some funding exists but is inconsistent.	One team gets AI budget.		
Resources	AI Budgeting & Funding	Level 1 – Initial	No dedicated AI funding.	AI projects rely on leftover budgets.		
			Total AI Budgeting/Funding		#DIV/0!	
Resources	AI Tools & Platforms	Level 5 – Optimized	Tooling is automated, scalable, and self-service across the org.	Auto-provisioned GPU clusters.		
Resources	AI Tools & Platforms	Level 4 – Managed	Standardized AI platforms are used across teams.	Enterprise MLOps platform.		
Resources	AI Tools & Platforms	Level 3 – Defined	Tools and platforms are documented and supported.	Approved AI tool catalog.		
Resources	AI Tools & Platforms	Level 2 – Developing	Tools exist but are inconsistent across teams.	One team uses Databricks, others don't.		
Resources	AI Tools & Platforms	Level 1 – Initial	No AI tooling strategy.	Teams use ad-hoc tools.		
			Total AI Tools/Platforms		#DIV/0!	

Data Security

Control Category	Control Name	Maturity Level	Description / Criteria	Example	Score	Comments
Data Security	Data Governance & Quality	Level 5 – Optimized	Automated data validation, lineage tracking, and quality scoring across all pipelines.	Real-time data quality scoring with auto-remediation.		
Data Security	Data Governance & Quality	Level 4 – Managed	Standardized data governance enforced across teams.	Mandatory data lineage tracking.		
Data Security	Data Governance & Quality	Level 3 – Defined	Documented data governance processes exist.	Data dictionaries, lineage diagrams.		
Data Security	Data Governance & Quality	Level 2 – Developing	Partial governance; inconsistent quality checks.	Some datasets validated manually.		
Data Security	Data Governance & Quality	Level 1 – Initial	No data governance or quality controls.	No validation of training data.		
				Total Data Governance/Quality	#DIV/0!	
Data Security	Data Poisoning Defense	Level 5 – Optimized	Autonomous detection and mitigation of poisoning attempts across all data sources.	Auto-flagging anomalous training samples.		
Data Security	Data Poisoning Defense	Level 4 – Managed	Continuous monitoring for poisoning indicators.	Drift detection tied to alerts.		
Data Security	Data Poisoning Defense	Level 3 – Defined	Automated validation and sanitization pipelines.	Outlier filtering before training.		
Data Security	Data Poisoning Defense	Level 2 – Developing	Manual dataset reviews.	Manual spot-checks for anomalies.		
Data Security	Data Poisoning Defense	Level 1 – Initial	No poisoning defenses.	Any data can enter training.		
				Total Data Poisoning Defense	#DIV/0!	

Data Security	Sensitive Data Protection	Level 5 – Optimized	Automated PII detection, redaction, and encryption across all AI workflows.	Auto-redaction of PII in prompts.		
Data Security	Sensitive Data Protection	Level 4 – Managed	Consistent enforcement of data protection policies.	PII scanning in pipelines.		
Data Security	Sensitive Data Protection	Level 3 – Defined	Documented data protection standards.	PII classification guidelines.		
Data Security	Sensitive Data Protection	Level 2 – Developing	Partial PII handling controls.	Manual redaction.		
Data Security	Sensitive Data Protection	Level 1 – Initial	No PII protection.	Sensitive data appears in training sets.		
				Total Sensitive Data Protection	#DIV/0!	

Model Security

Control Category	Control Name	Maturity Level	Description / Criteria	Example	Score	Comments
Model Security	Model Hardening	Level 5 – Optimized	Automated adversarial training and continuous robustness testing.	Auto-generated adversarial examples.		
Model Security	Model Hardening	Level 4 – Managed	Regular adversarial testing and model robustness reviews.	Quarterly red-team testing.		
Model Security	Model Hardening	Level 3 – Defined	Documented hardening techniques applied.	Gradient masking, noise injection.		
Model Security	Model Hardening	Level 2 – Developing	Some hardening applied inconsistently.	One model uses adversarial training.		
Model Security	Model Hardening	Level 1 – Initial	No model hardening.	Models vulnerable to simple attacks.		
			Total Model Hardening		#DIV/0!	
Model Security	Model Theft Defense	Level 5 – Optimized	Zero-Trust inference isolation with automated extraction detection.	Auto-blocking suspicious query patterns.		
Model Security	Model Theft Defense	Level 4 – Managed	Continuous monitoring for extraction attempts.	Rate-limit anomalies trigger alerts.		
Model Security	Model Theft Defense	Level 3 – Defined	Standardized protections implemented.	Query rate limiting, watermarking.		
Model Security	Model Theft Defense	Level 2 – Developing	Basic protections exist.	Simple throttling.		
Model Security	Model Theft Defense	Level 1 – Initial	No protections.	Unlimited inference queries.		
			Total Model Theft Defense		#DIV/0!	

Model Security	Model Supply Chain Security	Level 5 – Optimized	Autonomous dependency verification and continuous SBOM monitoring.	Auto-flagging compromised ML libraries.	
Model Security	Model Supply Chain Security	Level 4 – Managed	Continuous scanning of ML dependencies.	Automated CVE scanning.	
Model Security	Model Supply Chain Security	Level 3 – Defined	SBOM required for all models.	MLflow + SBOM integration.	
Model Security	Model Supply Chain Security	Level 2 – Developing	Partial dependency tracking.	Some libraries tracked manually.	
Model Security	Model Supply Chain Security	Level 1 – Initial	No supply chain controls.	Unknown ML dependencies.	
Total Model Supply Chain Security					#DIV/0!
Model Security	Model Access Control	Level 5 – Optimized	Zero-Trust access with continuous authentication and risk scoring.	Dynamic access revocation.	
Model Security	Model Access Control	Level 4 – Managed	Role-based access enforced across all models.	RBAC for inference endpoints.	
Model Security	Model Access Control	Level 3 – Defined	Access control policies documented.	Access matrix for models.	
Model Security	Model Access Control	Level 2 – Developing	Some access controls exist.	API keys per team.	
Model Security	Model Access Control	Level 1 – Initial	No access control.	Anyone can query models.	
Total Model Access Control					#DIV/0!

LLM Security

Control Category	Control Name	Maturity Level	Description / Criteria	Example	Score	Comments
LLM Security	Prompt Injection Defense	Level 5 – Optimized	Autonomous detection and blocking of direct/indirect prompt injection using AI-native Zero Trust.	System auto-isolates malicious prompt patterns.		
LLM Security	Prompt Injection Defense	Level 4 – Managed	ML-based detection of injection attempts with automated alerts.	Behavioral prompt anomaly detection.		
LLM Security	Prompt Injection Defense	Level 3 – Defined	Standardized prompt validation and sanitization pipeline.	Context filtering, allow/deny lists.		
LLM Security	Prompt Injection Defense	Level 2 – Developing	Basic keyword-based filtering.	Regex for “ignore previous instructions”.		
LLM Security	Prompt Injection Defense	Level 1 – Initial	No protections.	Model accepts any prompt.		
			Total Prompt Injection Defense		#DIV/0!	
LLM Security	Output Hardening	Level 5 – Optimized	Autonomous output filtering, redaction, and safety scoring.	Auto-blocking harmful outputs.		
LLM Security	Output Hardening	Level 4 – Managed	ML-based output moderation with continuous tuning.	Toxicity classifier + human review.		
LLM Security	Output Hardening	Level 3 – Defined	Standardized output validation rules.	Safety filters for PII, harmful content.		
LLM Security	Output Hardening	Level 2 – Developing	Basic output checks.	Simple profanity filter.		
LLM Security	Output Hardening	Level 1 – Initial	No output controls.	Model outputs unsafe content.		
			Total Output Hardening		#DIV/0!	
LLM Security	Hallucination Mitigation	Level 5 – Optimized	Automated fact-checking and retrieval-augmented validation.	RAG pipeline auto-corrects hallucinations.		

LLM Security	Hallucination Mitigation	Level 4 – Managed	Continuous monitoring of hallucination rates.	Hallucination KPI dashboard.	
LLM Security	Hallucination Mitigation	Level 3 – Defined	Documented hallucination controls.	Retrieval-augmented generation (RAG).	
LLM Security	Hallucination Mitigation	Level 2 – Developing	Some mitigation techniques used inconsistently.	Prompt engineering for accuracy.	
LLM Security	Hallucination Mitigation	Level 1 – Initial	No hallucination controls.	Model fabricates facts.	
Total Hallucination Mitigation					#DIV/0!
LLM Security	Data Leakage Prevention	Level 5 – Optimized	Autonomous detection and blocking of sensitive data in prompts and outputs.	Auto-redaction of PII in real time.	
LLM Security	Data Leakage Prevention	Level 4 – Managed	Continuous monitoring for leakage patterns.	Alerts for sensitive data exposure.	
LLM Security	Data Leakage Prevention	Level 3 – Defined	Standardized DLP rules for LLMs.	PII scanning in prompts.	
LLM Security	Data Leakage Prevention	Level 2 – Developing	Partial DLP controls.	Manual redaction.	
LLM Security	Data Leakage Prevention	Level 1 – Initial	No DLP controls.	Model leaks internal data.	
Total Data Leakage Prevention					#DIV/0!

Agent Security

Control Category	Control Name	Maturity Level	Description / Criteria	Example	Score	Comments
Agent Security	Tool-Use Security	Level 5 – Optimized	Autonomous risk scoring of tool calls with dynamic privilege adjustment.	Agent loses access if behavior becomes risky.		
Agent Security	Tool-Use Security	Level 4 – Managed	Continuous monitoring of tool usage.	Alerts for unusual tool invocation patterns.		
Agent Security	Tool-Use Security	Level 3 – Defined	Role-based tool access and guardrails.	Agent can only call approved tools.		
Agent Security	Tool-Use Security	Level 2 – Developing	Basic tool restrictions.	Hardcoded allow-list.		
Agent Security	Tool-Use Security	Level 1 – Initial	No tool-use controls.	Agent can execute any tool.		
			Total Tool Use Security		#DIV/0!	
Agent Security	Agent Memory Security	Level 5 – Optimized	Zero-Trust memory segmentation with autonomous poisoning detection.	Memory auto-cleans malicious entries.		
Agent Security	Agent Memory Security	Level 4 – Managed	Continuous monitoring of memory integrity.	Alerts for suspicious memory updates.		
Agent Security	Agent Memory Security	Level 3 – Defined	Standardized memory sanitization pipeline.	Validation before writing to memory.		
Agent Security	Agent Memory Security	Level 2 – Developing	Partial memory controls.	Manual memory review.		
Agent Security	Agent Memory Security	Level 1 – Initial	No memory protections.	Agent stores untrusted data.		
			Total Agent Memory Security		#DIV/0!	
Agent Security	Agent Autonomy Controls	Level 5 – Optimized	Autonomous risk-based throttling of agent autonomy.	Agent autonomy reduced during anomalies.		
Agent Security	Agent Autonomy Controls	Level 4 – Managed	Autonomy levels monitored and governed.	Autonomy KPI dashboard.		

Agent Security	Agent Autonomy Controls	Level 3 – Defined	Documented autonomy levels and policies.	Tiered autonomy model.	
Agent Security	Agent Autonomy Controls	Level 2 – Developing	Some autonomy restrictions.	Hardcoded limits.	
Agent Security	Agent Autonomy Controls	Level 1 – Initial	No autonomy controls.	Agent can act freely.	
			Total Agent Autonomy Controls		#DIV/0!
Agent Security	Agent Sandbox & Isolation	Level 5 – Optimized	Zero-Trust execution environment with autonomous isolation.	Agent auto-sandboxed during risky behavior.	
Agent Security	Agent Sandbox & Isolation	Level 4 – Managed	Continuous monitoring of agent execution.	Alerts for sandbox escapes.	
Agent Security	Agent Sandbox & Isolation	Level 3 – Defined	Standardized sandboxing for all agents.	Containerized agent execution.	
Agent Security	Agent Sandbox & Isolation	Level 2 – Developing	Partial sandboxing.	Some agents run in isolated VMs.	
Agent Security	Agent Sandbox & Isolation	Level 1 – Initial	No sandboxing.	Agents run with full system access.	
			Total Agent Sandbox Isolation		#DIV/0!

LLM/Agent Safety

Control Category	Control Name	Maturity Level	Description / Criteria	Example	Score	Comments
LLM/Agent Safety	Safety Guardrails	Level 5 – Optimized	Autonomous guardrail enforcement with dynamic policy adaptation.	Guardrails adjust based on context.		
LLM/Agent Safety	Safety Guardrails	Level 4 – Managed	Continuous monitoring of guardrail effectiveness.	Safety KPI dashboard.		
LLM/Agent Safety	Safety Guardrails	Level 3 – Defined	Standardized guardrails across all LLMs/agents.	Safety policies in prompt templates.		
LLM/Agent Safety	Safety Guardrails	Level 2 – Developing	Partial guardrails.	Hardcoded safety prompts.		
LLM/Agent Safety	Safety Guardrails	Level 1 – Initial	No guardrails.	Model outputs unsafe actions.		
			Total Safety Guardrails		#DIV/0!	

ATLAS Tactic Coverage

Control Category	Control Name	Maturity Level	Description / Criteria	Example	Score	Comments
ATLAS Tactic	Reconnaissance Defense	Level 5 – Optimized	Autonomous detection of reconnaissance patterns with automated blocking.	Auto-blocking repeated probing queries.		
ATLAS Tactic	Reconnaissance Defense	Level 4 – Managed	Continuous monitoring for reconnaissance attempts.	Alerts for unusual metadata requests.		
ATLAS Tactic	Reconnaissance Defense	Level 3 – Defined	Documented defenses against reconnaissance.	Rate limits, metadata suppression.		
ATLAS Tactic	Reconnaissance Defense	Level 2 – Developing	Partial protections.	Basic throttling.		
ATLAS Tactic	Reconnaissance Defense	Level 1 – Initial	No protections.	System reveals model details freely.		
			Total Reconnaissance Defense		#DIV/0!	
ATLAS Tactic	Initial Access Defense	Level 5 – Optimized	Zero-Trust access with continuous authentication and risk scoring.	Dynamic access revocation.		
ATLAS Tactic	Initial Access Defense	Level 4 – Managed	Strong authentication and monitoring.	MFA + behavioral analytics.		
ATLAS Tactic	Initial Access Defense	Level 3 – Defined	Documented access control policies.	RBAC for AI endpoints.		
ATLAS Tactic	Initial Access Defense	Level 2 – Developing	Basic access controls.	API keys.		
ATLAS Tactic	Initial Access Defense	Level 1 – Initial	No access controls.	Open inference endpoints.		
			Total Initial Access Defense		#DIV/0!	
ATLAS Tactic	Execution Defense	Level 5 – Optimized	Autonomous detection and containment of malicious execution.	Auto-sandboxing malicious actions.		
ATLAS Tactic	Execution Defense	Level 4 – Managed	Continuous monitoring of execution behavior.	Alerts for abnormal inference patterns.		

ATLAS Tactic	Execution Defense	Level 3 – Defined	Standardized execution controls.	Execution allow/deny lists.	
ATLAS Tactic	Execution Defense	Level 2 – Developing	Partial execution controls.	Some commands blocked.	
ATLAS Tactic	Execution Defense	Level 1 – Initial	No execution controls.	Any action allowed.	
			Total Execution Defense		#DIV/0!
ATLAS Tactic	Persistence Defense	Level 5 – Optimized	Autonomous detection of persistent footholds.	Auto-revocation of suspicious tokens.	
ATLAS Tactic	Persistence Defense	Level 4 – Managed	Continuous monitoring for persistence indicators.	Alerts for long-lived sessions.	
ATLAS Tactic	Persistence Defense	Level 3 – Defined	Documented persistence defenses.	Token rotation policies.	
ATLAS Tactic	Persistence Defense	Level 2 – Developing	Partial persistence controls.	Manual token cleanup.	
ATLAS Tactic	Persistence Defense	Level 1 – Initial	No persistence controls.	Tokens never expire.	
			Total Persistence Defense		#DIV/0!
ATLAS Tactic	Privilege Escalation Defense	Level 5 – Optimized	Autonomous privilege reduction based on risk.	Auto-downgrading suspicious accounts.	
ATLAS Tactic	Privilege Escalation Defense	Level 4 – Managed	Continuous monitoring of privilege changes.	Alerts for privilege anomalies.	
ATLAS Tactic	Privilege Escalation Defense	Level 3 – Defined	Standardized privilege policies.	RBAC + least privilege.	
ATLAS Tactic	Privilege Escalation Defense	Level 2 – Developing	Partial privilege controls.	Manual privilege reviews.	
ATLAS Tactic	Privilege Escalation Defense	Level 1 – Initial	No privilege controls.	All users have admin access.	
			Total Privilege Escalation Defense		#DIV/0!
ATLAS Tactic	Defense Evasion	Level 5 – Optimized	Autonomous detection of evasion patterns.	Auto-flagging obfuscated prompts.	

ATLAS Tactic	Defense Evasion	Level 4 – Managed	Continuous monitoring for evasion attempts.	Alerts for bypass attempts.	
ATLAS Tactic	Defense Evasion	Level 3 – Defined	Documented evasion defenses.	Prompt normalization.	
ATLAS Tactic	Defense Evasion	Level 2 – Developing	Partial defenses.	Basic pattern checks.	
ATLAS Tactic	Defense Evasion	Level 1 – Initial	No defenses.	Evasion attempts succeed.	
Total Defense Evasion					#DIV/0!
ATLAS Tactic	Credential Access	Level 5 – Optimized	Autonomous detection of credential theft attempts.	Auto-revocation of compromised keys.	
ATLAS Tactic	Credential Access	Level 4 – Managed	Continuous monitoring of credential usage.	Alerts for unusual API key use.	
ATLAS Tactic	Credential Access	Level 3 – Defined	Standardized credential security.	Key rotation, vaulting.	
ATLAS Tactic	Credential Access	Level 2 – Developing	Partial credential controls.	Manual key rotation.	
ATLAS Tactic	Credential Access	Level 1 – Initial	No credential protections.	Keys stored in plaintext.	
Total Credential Access					#DIV/0!
ATLAS Tactic	Discovery Defense	Level 5 – Optimized	Autonomous detection of discovery attempts.	Auto-blocking enumeration patterns.	
ATLAS Tactic	Discovery Defense	Level 4 – Managed	Continuous monitoring for discovery behavior.	Alerts for probing queries.	
ATLAS Tactic	Discovery Defense	Level 3 – Defined	Documented discovery defenses.	Suppression of system metadata.	
ATLAS Tactic	Discovery Defense	Level 2 – Developing	Partial defenses.	Basic metadata hiding.	
ATLAS Tactic	Discovery Defense	Level 1 – Initial	No defenses.	System reveals architecture details.	
Total Discovery Defense					#DIV/0!
ATLAS Tactic	Lateral Movement	Level 5 – Optimized	Autonomous segmentation and isolation of AI components.	Auto-blocking cross-system access.	
ATLAS Tactic	Lateral Movement	Level 4 – Managed	Continuous monitoring for lateral movement.	Alerts for cross-service anomalies.	

ATLAS Tactic	Lateral Movement	Level 3 – Defined	Segmented AI architecture.	Separate inference and training networks.	
ATLAS Tactic	Lateral Movement	Level 2 – Developing	Partial segmentation.	Some network boundaries.	
ATLAS Tactic	Lateral Movement	Level 1 – Initial	No segmentation.	Flat network.	
Total Lateral Movement					#DIV/0!
ATLAS Tactic	Collection Defense	Level 5 – Optimized	Autonomous detection of data exfiltration patterns.	Auto-blocking large data pulls.	
ATLAS Tactic	Collection Defense	Level 4 – Managed	Continuous monitoring for data collection anomalies.	Alerts for bulk inference queries.	
ATLAS Tactic	Collection Defense	Level 3 – Defined	Documented data collection controls.	Query rate limits.	
ATLAS Tactic	Collection Defense	Level 2 – Developing	Partial controls.	Manual review of logs.	
ATLAS Tactic	Collection Defense	Level 1 – Initial	No controls.	Unlimited data extraction.	
Total Collection Defense					#DIV/0!
ATLAS Tactic	Exfiltration Defense	Level 5 – Optimized	Autonomous detection and blocking of exfiltration attempts.	Auto-blocking suspicious downloads.	
ATLAS Tactic	Exfiltration Defense	Level 4 – Managed	Continuous monitoring for exfiltration patterns.	Alerts for large model exports.	
ATLAS Tactic	Exfiltration Defense	Level 3 – Defined	Standardized exfiltration controls.	Encryption, access logs.	
ATLAS Tactic	Exfiltration Defense	Level 2 – Developing	Partial controls.	Manual approval for exports.	
ATLAS Tactic	Exfiltration Defense	Level 1 – Initial	No controls.	Anyone can export models.	
Total Exfiltration Defense					#DIV/0!
ATLAS Tactic	Impact Defense	Level 5 – Optimized	Autonomous mitigation of harmful impacts.	Auto-rollback after harmful outputs.	
ATLAS Tactic	Impact Defense	Level 4 – Managed	Continuous monitoring of impact indicators.	Alerts for harmful model behavior.	
ATLAS Tactic	Impact Defense	Level 3 – Defined	Documented impact mitigation processes.	Safety guardrails.	

ATLAS Tactic	Impact Defense	Level 2 – Developing	Partial impact controls.	Manual review of harmful outputs.	
ATLAS Tactic	Impact Defense	Level 1 – Initial	No impact controls.	Harmful outputs go unnoticed.	
Total Impact Defense					#DIV/0!
ATLAS Tactic	Model Manipulation Defense	Level 5 – Optimized	Autonomous detection of manipulation attempts.	Auto-flagging adversarial prompts.	
ATLAS Tactic	Model Manipulation Defense	Level 4 – Managed	Continuous monitoring for manipulation.	Alerts for adversarial inputs.	
ATLAS Tactic	Model Manipulation Defense	Level 3 – Defined	Documented manipulation defenses.	Adversarial training.	
ATLAS Tactic	Model Manipulation Defense	Level 2 – Developing	Partial defenses.	Basic input validation.	
ATLAS Tactic	Model Manipulation Defense	Level 1 – Initial	No defenses.	Model easily manipulated.	
Total Model Manipulation Defense					#DIV/0!
ATLAS Tactic	Model Degradation Defense	Level 5 – Optimized	Autonomous detection and mitigation of degradation.	Auto-retraining after degradation.	
ATLAS Tactic	Model Degradation Defense	Level 4 – Managed	Continuous monitoring of degradation indicators.	Drift + performance dashboards.	
ATLAS Tactic	Model Degradation Defense	Level 3 – Defined	Documented degradation controls.	Performance thresholds.	
ATLAS Tactic	Model Degradation Defense	Level 2 – Developing	Partial controls.	Manual performance checks.	
ATLAS Tactic	Model Degradation Defense	Level 1 – Initial	No controls.	Degradation unnoticed.	
Total Model Degradation Defense					#DIV/0!

ATLAS Mitigation

Control Category	Control Name	Maturity Level	Description / Criteria	Example	Score	Comments
ATLAS Mitigation	Adversarial Training	Level 5 – Optimized	Continuous adversarial retraining with automated generation of adversarial examples.	Auto-generated adversarial samples.		
ATLAS Mitigation	Adversarial Training	Level 4 – Managed	Regular adversarial testing and retraining.	Quarterly adversarial robustness tests.		
ATLAS Mitigation	Adversarial Training	Level 3 – Defined	Documented adversarial training procedures.	FGSM, PGD training.		
ATLAS Mitigation	Adversarial Training	Level 2 – Developing	Partial adversarial training.	One model hardened.		
ATLAS Mitigation	Adversarial Training	Level 1 – Initial	No adversarial training.	Model vulnerable to simple attacks.		
			Total Adversarial Training		#DIV/0!	
ATLAS Mitigation	Input Validation	Level 5 – Optimized	Autonomous input sanitization with AI-native Zero Trust.	Auto-blocking malicious prompts.		
ATLAS Mitigation	Input Validation	Level 4 – Managed	ML-based input anomaly detection.	Behavioral input classifier.		
ATLAS Mitigation	Input Validation	Level 3 – Defined	Standardized input validation rules.	Allow/deny lists.		
ATLAS Mitigation	Input Validation	Level 2 – Developing	Basic validation.	Regex filtering.		
ATLAS Mitigation	Input Validation	Level 1 – Initial	No validation.	Any input accepted.		
			Total Input Validation		#DIV/0!	
ATLAS Mitigation	Data Sanitization	Level 5 – Optimized	Fully automated sanitization with anomaly detection and poisoning defense.	Auto-flagging poisoned samples.		

ATLAS Mitigation	Data Sanitization	Level 4 – Managed	Continuous monitoring of data quality.	Drift + anomaly alerts.	
ATLAS Mitigation	Data Sanitization	Level 3 – Defined	Standardized sanitization pipeline.	Outlier removal, deduplication.	
ATLAS Mitigation	Data Sanitization	Level 2 – Developing	Partial sanitization.	Manual dataset review.	
ATLAS Mitigation	Data Sanitization	Level 1 – Initial	No sanitization.	Raw data used directly.	
Total Data Sanitization					#DIV/0!
ATLAS Mitigation	Data Label Verification	Level 5 – Optimized	Automated label validation using ML + human-in-the-loop escalation.	Auto-flag mislabeled samples.	
ATLAS Mitigation	Data Label Verification	Level 4 – Managed	Continuous monitoring of label integrity.	Label drift alerts.	
ATLAS Mitigation	Data Label Verification	Level 3 – Defined	Standardized label QA processes.	Dual-annotator workflow.	
ATLAS Mitigation	Data Label Verification	Level 2 – Developing	Partial label checks.	Spot-checking labels.	
ATLAS Mitigation	Data Label Verification	Level 1 – Initial	No label verification.	Labels taken at face value.	
Total Data Label Verification					#DIV/0!
ATLAS Mitigation	Data Provenance Tracking	Level 5 – Optimized	Automated provenance tracking with immutable lineage.	Blockchain-backed lineage.	
ATLAS Mitigation	Data Provenance Tracking	Level 4 – Managed	Continuous monitoring of data sources.	Alerts for untrusted sources.	
ATLAS Mitigation	Data Provenance Tracking	Level 3 – Defined	Documented provenance requirements.	Source metadata stored in registry.	
ATLAS Mitigation	Data Provenance Tracking	Level 2 – Developing	Partial provenance tracking.	Some datasets tagged manually.	
ATLAS Mitigation	Data Provenance Tracking	Level 1 – Initial	No provenance tracking.	Unknown data origins.	
Total Data Provenance Tracking					#DIV/0!

ATLAS Mitigation	Secure Data Storage	Level 5 – Optimized	Autonomous encryption, key rotation, and access scoring.	Auto-revoking risky access.	
ATLAS Mitigation	Secure Data Storage	Level 4 – Managed	Continuous monitoring of data access.	Alerts for unusual reads.	
ATLAS Mitigation	Secure Data Storage	Level 3 – Defined	Standardized encryption and access policies.	KMS + RBAC.	
ATLAS Mitigation	Secure Data Storage	Level 2 – Developing	Partial encryption.	Some datasets encrypted.	
ATLAS Mitigation	Secure Data Storage	Level 1 – Initial	No secure storage.	Plaintext training data.	
Total AI Threat Analytics					#DIV/0!
ATLAS Mitigation	Training Pipeline Integrity	Level 5 – Optimized	Autonomous integrity verification with rollback.	Auto-rollback on pipeline tampering.	
ATLAS Mitigation	Training Pipeline Integrity	Level 4 – Managed	Continuous monitoring of pipeline integrity.	Alerts for unauthorized changes.	
ATLAS Mitigation	Training Pipeline Integrity	Level 3 – Defined	Standardized pipeline security controls.	Signed training jobs.	
ATLAS Mitigation	Training Pipeline Integrity	Level 2 – Developing	Partial controls.	Manual pipeline checks.	
ATLAS Mitigation	Training Pipeline Integrity	Level 1 – Initial	No integrity controls.	Anyone can modify training code.	
Total Training Pipeline Integrity					#DIV/0!
ATLAS Mitigation	Model Validation & Testing	Level 5 – Optimized	Autonomous validation with safety, robustness, and fairness scoring.	Auto-blocking unsafe models.	
ATLAS Mitigation	Model Validation & Testing	Level 4 – Managed	Continuous validation with dashboards.	Safety + robustness KPIs.	
ATLAS Mitigation	Model Validation & Testing	Level 3 – Defined	Standardized validation suite.	Unit tests, adversarial tests.	

ATLAS Mitigation	Model Validation & Testing	Level 2 – Developing	Partial validation.	Manual testing.	
ATLAS Mitigation	Model Validation & Testing	Level 1 – Initial	No validation.	Models deployed untested.	
Total Model Validation/Testing					#DIV/0!
ATLAS Mitigation	Model Hardening	Level 5 – Optimized	Automated adversarial robustness testing and continuous hardening.	Auto-generated adversarial examples.	
ATLAS Mitigation	Model Hardening	Level 4 – Managed	Regular robustness evaluations.	Quarterly adversarial testing.	
ATLAS Mitigation	Model Hardening	Level 3 – Defined	Documented hardening techniques.	Gradient masking, noise injection.	
ATLAS Mitigation	Model Hardening	Level 2 – Developing	Partial hardening.	One model hardened.	
ATLAS Mitigation	Model Hardening	Level 1 – Initial	No hardening.	Model vulnerable to simple attacks.	
Total Model Hardening					#DIV/0!
ATLAS Mitigation	Model Watermarking	Level 5 – Optimized	Automated watermarking with tamper-resistant verification.	Watermark survives adversarial removal.	
ATLAS Mitigation	Model Watermarking	Level 4 – Managed	Continuous monitoring for watermark integrity.	Alerts for watermark anomalies.	
ATLAS Mitigation	Model Watermarking	Level 3 – Defined	Standardized watermarking process.	Embedded signature in model weights.	
ATLAS Mitigation	Model Watermarking	Level 2 – Developing	Partial watermarking.	Only some models watermarked.	
ATLAS Mitigation	Model Watermarking	Level 1 – Initial	No watermarking.	No way to detect model theft.	
Total Model Watermarking					#DIV/0!
ATLAS Mitigation	Model Explainability	Level 5 – Optimized	Automated explainability scoring and continuous transparency checks.	Auto-flagging unexplained decisions.	
ATLAS Mitigation	Model Explainability	Level 4 – Managed	Continuous monitoring of explainability metrics.	Explainability dashboards.	

ATLAS Mitigation	Model Explainability	Level 3 – Defined	Standardized explainability methods.	SHAP, LIME, attention maps.	
ATLAS Mitigation	Model Explainability	Level 2 – Developing	Partial explainability.	Some models have explanations.	
ATLAS Mitigation	Model Explainability	Level 1 – Initial	No explainability.	Black-box models with no insight.	
			Total Model Explainability		#DIV/0!
ATLAS Mitigation	Model Robustness Testing	Level 5 – Optimized	Autonomous robustness testing with continuous updates.	Auto-generated stress tests.	
ATLAS Mitigation	Model Robustness Testing	Level 4 – Managed	Regular robustness evaluations.	Quarterly robustness reports.	
ATLAS Mitigation	Model Robustness Testing	Level 3 – Defined	Standardized robustness testing suite.	Noise, perturbation, adversarial tests.	
ATLAS Mitigation	Model Robustness Testing	Level 2 – Developing	Partial robustness testing.	Manual stress tests.	
ATLAS Mitigation	Model Robustness Testing	Level 1 – Initial	No robustness testing.	Model fails under simple perturbations.	
			Total Model Robustness Testing		#DIV/0!
ATLAS Mitigation	Model Access Control	Level 5 – Optimized	Zero-Trust access with continuous risk scoring.	Dynamic access revocation.	
ATLAS Mitigation	Model Access Control	Level 4 – Managed	Continuous monitoring of access patterns.	Alerts for unusual inference usage.	
ATLAS Mitigation	Model Access Control	Level 3 – Defined	Standardized RBAC/ABAC for models.	Role-based inference access.	
ATLAS Mitigation	Model Access Control	Level 2 – Developing	Partial access controls.	API keys per team.	
ATLAS Mitigation	Model Access Control	Level 1 – Initial	No access control.	Anyone can query models.	
			Total Model Access Control		#DIV/0!
ATLAS Mitigation	Inference Rate Limiting	Level 5 – Optimized	Adaptive rate limiting based on risk scoring.	Auto-throttling suspicious users.	

ATLAS Mitigation	Inference Rate Limiting	Level 4 – Managed	Continuous monitoring of inference volume.	Alerts for extraction-like patterns.	
ATLAS Mitigation	Inference Rate Limiting	Level 3 – Defined	Standardized rate limits.	Per-user/per-IP quotas.	
ATLAS Mitigation	Inference Rate Limiting	Level 2 – Developing	Basic rate limits.	Simple per-minute caps.	
ATLAS Mitigation	Inference Rate Limiting	Level 1 – Initial	No rate limits.	Unlimited queries.	
			Total Interference Rate Limiting		#DIV/0!
ATLAS Mitigation	Output Filtering	Level 5 – Optimized	Autonomous output moderation with dynamic policy adaptation.	Auto-blocking harmful outputs.	
ATLAS Mitigation	Output Filtering	Level 4 – Managed	Continuous tuning of moderation models.	Toxicity classifier + human review.	
ATLAS Mitigation	Output Filtering	Level 3 – Defined	Standardized output filtering rules.	PII redaction, safety filters.	
ATLAS Mitigation	Output Filtering	Level 2 – Developing	Basic filtering.	Profanity filter.	
ATLAS Mitigation	Output Filtering	Level 1 – Initial	No filtering.	Model outputs unsafe content.	
			Total Output Filtering		#DIV/0!
ATLAS Mitigation	Secure Infrastructure Configuration	Level 5 – Optimized	Autonomous configuration scanning and auto-remediation.	Auto-patching GPU nodes.	
ATLAS Mitigation	Secure Infrastructure Configuration	Level 4 – Managed	Continuous monitoring of infra security posture.	Alerts for insecure configs.	
ATLAS Mitigation	Secure Infrastructure Configuration	Level 3 – Defined	Standardized hardening baselines.	CIS-aligned ML infra baselines.	
ATLAS Mitigation	Secure Infrastructure Configuration	Level 2 – Developing	Partial hardening.	Some nodes hardened.	
ATLAS Mitigation	Secure Infrastructure Configuration	Level 1 – Initial	No hardening.	Default insecure infra.	
			Total Secure Infrastructure Configuration		#DIV/0!

ATLAS Mitigation	Network Segmentation	Level 5 – Optimized	Dynamic segmentation with autonomous isolation.	Auto-isolating compromised nodes.	
ATLAS Mitigation	Network Segmentation	Level 4 – Managed	Continuous monitoring of network boundaries.	Alerts for cross-segment anomalies.	
ATLAS Mitigation	Network Segmentation	Level 3 – Defined	Standardized segmentation architecture.	Separate training/inference networks.	
ATLAS Mitigation	Network Segmentation	Level 2 – Developing	Partial segmentation.	Some VLANs or VPCs.	
ATLAS Mitigation	Network Segmentation	Level 1 – Initial	No segmentation.	Flat network.	
			Total Network Segmentation		#DIV/0!
ATLAS Mitigation	Runtime Monitoring	Level 5 – Optimized	Autonomous runtime anomaly detection with auto-containment.	Auto-sandboxing suspicious inference.	
ATLAS Mitigation	Runtime Monitoring	Level 4 – Managed	Continuous runtime monitoring.	Alerts for abnormal outputs.	
ATLAS Mitigation	Runtime Monitoring	Level 3 – Defined	Standardized runtime monitoring rules.	Output scoring, drift detection.	
ATLAS Mitigation	Runtime Monitoring	Level 2 – Developing	Partial runtime monitoring.	Logs only some inference events.	
ATLAS Mitigation	Runtime Monitoring	Level 1 – Initial	No runtime monitoring.	No visibility into inference behavior.	
			Total Runtime Monitoring		#DIV/0!
ATLAS Mitigation	Secure API Gateway	Level 5 – Optimized	AI-aware API gateway with adaptive threat scoring.	Auto-blocking malicious prompt patterns.	
ATLAS Mitigation	Secure API Gateway	Level 4 – Managed	Continuous monitoring of API traffic.	Alerts for extraction-like behavior.	
ATLAS Mitigation	Secure API Gateway	Level 3 – Defined	Standardized API security controls.	AuthN, AuthZ, rate limits.	
ATLAS Mitigation	Secure API Gateway	Level 2 – Developing	Basic API protections.	API keys + simple throttling.	
ATLAS Mitigation	Secure API Gateway	Level 1 – Initial	No API gateway.	Direct access to model endpoints.	

Total Secure API Gateway					#DIV/0!
ATLAS Mitigation	Key & Secret Management	Level 5 – Optimized	Autonomous key rotation and compromise detection.	Auto-revoking leaked keys.	
ATLAS Mitigation	Key & Secret Management	Level 4 – Managed	Continuous monitoring of key usage.	Alerts for unusual key activity.	
ATLAS Mitigation	Key & Secret Management	Level 3 – Defined	Standardized key vaulting and rotation.	KMS + HSM.	
ATLAS Mitigation	Key & Secret Management	Level 2 – Developing	Partial key management.	Manual rotation.	
ATLAS Mitigation	Key & Secret Management	Level 1 – Initial	No key management.	Keys stored in plaintext.	
Total Key/Secret Management					#DIV/0!
ATLAS Mitigation	Logging & Audit Trails	Level 5 – Optimized	Immutable logs with autonomous anomaly detection.	Blockchain-backed audit logs.	
ATLAS Mitigation	Logging & Audit Trails	Level 4 – Managed	Continuous log monitoring.	Alerts for suspicious activity.	
ATLAS Mitigation	Logging & Audit Trails	Level 3 – Defined	Standardized logging across AI systems.	Input/output logging.	
ATLAS Mitigation	Logging & Audit Trails	Level 2 – Developing	Partial logging.	Logs only some events.	
ATLAS Mitigation	Logging & Audit Trails	Level 1 – Initial	No logging.	No traceability.	
Total Logging/Audit Trails					#DIV/0!
ATLAS Mitigation	Supply Chain Security	Level 5 – Optimized	Autonomous SBOM scanning + dependency integrity verification.	Auto-flagging compromised ML libs.	
ATLAS Mitigation	Supply Chain Security	Level 4 – Managed	Continuous scanning of dependencies.	CVE alerts for ML frameworks.	
ATLAS Mitigation	Supply Chain Security	Level 3 – Defined	SBOM required for all models.	Signed dependencies.	
ATLAS Mitigation	Supply Chain Security	Level 2 – Developing	Partial supply chain controls.	Manual dependency checks.	
ATLAS Mitigation	Supply Chain Security	Level 1 – Initial	No supply chain security.	Unknown dependencies.	
Total Supply Chain Security					#DIV/0!

ATLAS Mitigation	AI Incident Response	Level 5 – Optimized	Autonomous containment + automated recovery workflows.	Auto-quarantine of compromised model.	
ATLAS Mitigation	AI Incident Response	Level 4 – Managed	Continuous monitoring + rehearsed IR plans.	AI red-team exercises.	
ATLAS Mitigation	AI Incident Response	Level 3 – Defined	Documented AI-specific IR playbooks.	ATLAS-aligned IR procedures.	
ATLAS Mitigation	AI Incident Response	Level 2 – Developing	Partial IR processes.	Manual response to AI failures.	
ATLAS Mitigation	AI Incident Response	Level 1 – Initial	No AI-specific IR.	No plan for LLM compromise.	
			Total AI Incident Response		#DIV/0!
ATLAS Mitigation	Secure Deployment Pipeline	Level 5 – Optimized	Autonomous validation + Zero Trust enforcement.	Auto-blocking unsafe deployments.	
ATLAS Mitigation	Secure Deployment Pipeline	Level 4 – Managed	Continuous scanning of deployment artifacts.	Alerts for tampered models.	
ATLAS Mitigation	Secure Deployment Pipeline	Level 3 – Defined	Standardized deployment security.	Signed model artifacts.	
ATLAS Mitigation	Secure Deployment Pipeline	Level 2 – Developing	Partial deployment controls.	Manual artifact checks.	
ATLAS Mitigation	Secure Deployment Pipeline	Level 1 – Initial	No deployment security.	Anyone can deploy anything.	
			Total Secure Deployment Pipeline		#DIV/0!

Shadow AI Detection Framework (SAiDF)

A behavioral methodology for detecting AI-mediated trust abuse.
What defenders use when malware never appears.

Traditional detection relies on malware artifacts. SAiDF focuses on behavioral indicators when compromise occurs entirely inside trusted workflows.

SAiDF focuses on:

- behavioral cadence
- workflow mediation
- prompt movement
- trust-boundary drift
- browser-native exfiltration

SAiDF prioritizes behavior over malware artifacts

Shadow AI Detection Framework (SAIDF) - Identity Layer

Detect:

- delegated AI identity usage
- token reuse across workflows
- non-human interaction patterns
- identity-origin drift

Core Signal

`identity_used ≠ identity_origin`

AI inherits identity trust without inheriting human accountability.

The authenticated identity is not always the entity performing the workflow

Shadow AI Detection Framework (SAIDF) – Behavior Layer

Detect:

- low-variance timing
- high-frequency calls
- deterministic intervals
- repetitive prompt structures

Core Signal

```
stddev(interval) < threshold
AND request_count > baseline × N
```

Human behavior contains variance.
Automated workflow mediation often does not.

Shadow AI Detection Framework (SAIDF) – Data Layer

Detect:

- structured prompts
- sensitive content movement
- source code / secret exposure
- Large payloads

Core Signal

```
payload_size > threshold
AND contains_sensitive_data = true
```

The risk is not only where data goes — but how prompts transform and expose it.

Shadow AI Detection Framework (SAIDF) – Execution Layer

Detect:

- autonomous task execution
- browser-mediated workflows
- delegated AI action and chained AI-driven actions

Core Signal

```
ai_action_count > threshold
AND user_interaction = false
```

Trusted workflows can now perform autonomous actions without meaningful human verification.

ShadowAI Detection Signals

Identity Signals

Detect delegated or non-human AI usage behavior
service-account AI interactions
multi-origin authentication drift
OAuth delegation anomalies
AI-linked identity reuse

```
index=auth_logs
| stats dc(src_ip) by user
| where user_type="service" AND dc(src_ip) > 5
```

Workflow Behavioral Signals

Detect machine-mediated AI interaction patterns
high-frequency prompt submission
low interaction variance
repetitive workflow structure
consistent payload cadence

```
index=proxy_logs
(dest_domain="api.openai.com" OR dest_domain="api.anthropic.com")
| stats count avg(bytes_out) stdev(_time) as variance by user
| where count > 100 AND variance < 2
```

Data Signals

Detect structured or sensitive prompt movement
high-entropy prompt payloads
abnormal POST sizes
repeated sensitive submissions
prompt exfiltration indicators

```
index=proxy_logs method=POST
| where bytes_out > 5000
| eval entropy=...
| where entropy > 4
```

Execution / Infrastructure Signals

Detect unmanaged AI execution paths
unsanctioned AI domains
shadow SaaS usage
unapproved AI endpoints
browser-mediated AI traffic

```
index=dns_logs
| search query IN ("openai.com","anthropic.com")
| stats count by user
| where user NOT IN approved_list
```

KPIs and Indicators

We define a measurement system that exposes AI-mediated activity across identity, data, and execution layers.
Operational visibility requires measurable behavioral telemetry.

A. Exposure (what exists)

- AI traffic (sanctioned vs unsanctioned)
- Unknown AI SaaS usage ratio
- Browser-extension AI mediation rate

B. Behavior (what users do)

- Shadow AI incidents per quarter
- Repeated structured AI interactions
- Prompt-Level DLP hit rate

C. Security effectiveness (what security catches)

- MTTD for Shadow AI activity
- MTTR for containment
- CASB enforcement rate
- AI prompt DLP detection rate
- Red-team detection coverage

D. Governance maturity (how controlled it is)

- Policy exception growth rate
- AI workflow visibility coverage
- Governance automation level

E. Business risk outcome (what it causes)

- Sensitive data leakage reduction
- Incident severity trend
- Shadow AI Exposure Index

We cannot secure what we do not measure — and ShadowAI requires a new measurement model.

KPIs and Indicators

How organizations measure Shadow AI exposure

A. Exposure (what exists)

- **AI traffic (sanctioned vs unsanctioned)**

How much AI usage is outside approved tools

Formula:

$$AI\ Traffic\ Ratio = \frac{Unsanctioned\ AI\ Traffic}{Total\ AI\ Traffic} * 100$$

implement

You need 3 data sources:

Proxy logs (Zscaler / Netskope / firewall)
CASB logs (Microsoft Defender, Netskope, etc.)
DNS logs

Classification logic:

Sanctioned = Copilot, enterprise AI tools
Unsanctioned = unknown LLM endpoints, shadow SaaS, browser-based tools

Output (what it looks like)

Month	Total AI Traffic	Unsanctioned	Ratio
Jan	1,000,000	300,000	30%

KPIs and Indicators

How organizations measure Shadow AI exposure

A. Exposure (what exists)

- **Unknown AI SaaS ratio**

How much AI usage is happening on services we have NOT classified or sanctioned

Formula: $$Unknown\ AI\ SaaS\ ratio = \frac{Unknown\ AI\ Domains}{Total\ AI\ Domains} * 100$$

Implement

Step 1 — Define AI traffic universe

You need to classify traffic as "AI-related" first.

DNS logs Proxy logs
CASB logs SASE logs (Zscaler / Netskope / Prisma Access)

Filter by:

known LLM keywords (chat, gpt, ai, assistant, copilot)
known AI endpoints (OpenAI, Anthropic, Google Gemini, etc.)
embedding APIs / inference APIs

Step 2 — Classify domains

Category	Meaning
Sanctioned AI SaaS	approved enterprise tools
Known unsanctioned AI SaaS	known public AI tools not approved
Unknown AI SaaS	not in any registry, but behaving like AI service

Step 3 — Identify "unknown AI SaaS"

This is the important part most models miss.

Flag as "unknown AI SaaS" if:

- domain not in allowlist/denylist
- API-like behavior detected (POST / completions, /chat, /generate)
- high entropy payloads (prompt-like strings)
- OpenAI-compatible schema patterns (JSON prompts, system/user roles)
- high text-to-response ratio endpoints

Output (what it looks like)

Month	AI Domains	Unknown AI SaaS	Ratio
Jan	120	18	15%
Feb	150	30	20%

Why this KPI is powerful

It helps identify:
previously unknown AI services
unmanaged AI-enabled SaaS platforms
indirect or embedded LLM usage inside enterprise workflows
emerging ShadowAI activity before formal governance exists
This becomes an operational discovery metric for AI exposure.

KPIs and Indicators

How organizations measure Shadow AI exposure

A. Exposure (what exists)

- **Browser-extension AI mediation rate**

How much AI usage is happening through browser extensions instead of sanctioned enterprise channels

Formula:

$$Browser\ Extension\ AI\ Mediation\ Rate = \frac{AI\ Sessions\ Using\ Extensions}{Total\ AI\ Sessions} * 100$$

Implement

Step 1 — Identify AI-capable extensions

You need inventory from:
- Chrome Enterprise Admin logs
- Edge extension management
- Endpoint agents (EDR like CrowdStrike, SentinelOne)
- Browser telemetry APIs (if available)

Classify extensions:
- ChatGPT helper extensions
- Grammarly / writing assistants
- "AI sidebar tools"
- code assistant plugins
- prompt injection tools

Step 2 — Detect AI traffic from extensions

Signals:
- HTTP user-agent includes extension ID
- request originates from chrome-extension://
- local proxy injection headers
- browser process → API calls not routed through enterprise proxy

Step 3 — Identify AI-bound requests

Same filtering logic:
- LLM endpoints
- generative API patterns
- prompt/response payload structures

Output

Month	Total AI Requests	Extension Requests	Rate
Jan	1,000,000	120,000	12%
Feb	1,200,000	180,000	15%

KPIs and Indicators

How organizations measure Shadow AI exposure

B. Behavior (what users do)

- **Shadow AI incidents per quarter**

Detected risky AI interactions requiring security action

Formula:

$$ShadowAI\ Incident\ per\ quarter = \frac{Total\ ShadowAI\ Incidents}{Quarter}$$

Implement

Deduplicate incidents:
Same user + same domain + same payload = 1 incident

Output

Quarter	Incidents	Trend
Q1	420	↑
Q2	310	↓

B. Behavior (what users do)

- **Repeated Structured AI Interactions**

Are users repeatedly using structured prompt patterns (automation behavior)?

Formula:

$$Repeated\ Structured\ AI\ Interactions = \frac{Repeated\ AI\ Sessions}{Total\ AI\ Sessions} * 100$$

Implement

A structured AI interaction = prompts with:
templates
repeated schema-like formatting
JSON prompts
workflow chaining ("step 1 → step 2 → step 3" patterns)

Output

Month	Users	Repeated Pattern Users	Rate
Jan	8,000	1,200	15%

KPIs and Indicators

How organizations measure Shadow AI exposure

B. Behavior (what users do)

- **Prompt DLP Hit Rate**

% of AI prompts containing sensitive data

Formula:

$$Prompt\ DLP\ Hit\ Rate = \frac{Prompt\ DLP\ Matches}{Total\ AI\ Prompts} * 100$$

Implement

Detection sources
- endpoint agent (browser / EDR)
- proxy payload inspection
- CASB inline inspection

Sensitive indicators
- PII (emails, SSNs)
- API keys
- internal code snippets
- confidential docs pasted into prompts

Output

Month	Sensitive Prompts	Total Prompts	Rate
Jan	3,400	80,000	4.25%

KPIs and Indicators

How organizations measure Shadow AI exposure

C. Security effectiveness (what security catches)

- **MTTD (Mean Time to Detect Shadow AI)**

How fast you detect shadow AI usage after it starts

Formula:

$$MTTD = \frac{Total\ Detection\ Time}{Number\ of\ Incidents}$$

Implement

First network request timestamp
Detection event timestamp (CASB/DLP/SIEM)

Output

Month	MTTD
Jan	48 min
Feb	22 min

C. Security effectiveness (what security catches)

- **MTTR (Mean Time to Respond)**

Time from Shadow AI detection → containment action completed.

Formula:

$$MTTR = \frac{Total\ Response\ Time}{Number\ of\ Incidents}$$

Implement

Field	Source
detection_time	CASB / SIEM
containment_time	firewall / proxy / IAM
action_type	block, revoke, isolate, warn
user_id	identity provider

Output

Incident	Detect Time	Contain Time	MTTR
#101	10:02	10:18	16 min
#102	14:05	14:09	4 min

KPIs and Indicators

How organizations measure Shadow AI exposure

C. Security effectiveness (what security catches)

- **AI Prompt DLP Hit Rate**

% of AI prompts containing sensitive data

Formula:

$$AI\ Prompt\ DLP\ Hit\ Rate = \frac{Sensitive\ Prompt\ Matches}{Total\ AI\ Uploads} * 100$$

Implement

How to detect prompts:
- browser extension logging (if allowed)
- endpoint agent (EDR)
- proxy inspection (if decrypted traffic)

Output

Month	Prompt Hits	Total Prompts	Rate
Jan	2,100	50,000	4.2%

C. Security effectiveness (what security catches)

- **CASB Blocks (Shadow AI enforcement KPI)**

Number of AI-related sessions blocked by CASB

Formula:

$$CASB\ Blocks = \frac{Blocked\ AI\ Sessions}{Total\ AI\ Sessions} * 100$$

Implement

Data Sources
- CASB logs
- Proxy / SWG telemetry
- Identity provider events
- Firewall enforcement logs

Detection Logic

Identify:
- blocked AI SaaS access
- denied OAuth grants
- blocked AI uploads
- revoked AI sessions
- AI policy violations

Containment Actions
- block AI domain
- revoke OAuth token
- isolate user session
- terminate browser session
- enforce step-up authentication

Output

Month	AI Sessions	Blocked	Rate
Jan	1.2M	240K	20%

KPIs and Indicators

How organizations measure Shadow AI exposure

C. Security effectiveness (what security catches)

- **Red-Team Findings (Shadow AI validation KPI)**

Security team simulated Shadow AI usage and detected exposures.

Formula:

$$Red\ Team\ Findings\ Rate = \frac{Detected\ Red-Team\ Activities}{Total\ Simulated\ Attacks} * 100$$

Implement

Simulated Scenarios

prompt injection
sensitive data leakage
browser-extension exfiltration
malicious AI plugin behavior
AI-assisted phishing
unauthorized AI SaaS usage
delegated OAuth abuse

Detection Sources

Prompt Visibility
browser extension logging (if permitted)
endpoint telemetry / EDR
proxy inspection (when TLS decryption exists)
CASB AI session monitoring
browser DevTools/network inspection during testing

Detection Objectives

Validate whether controls can detect:
prompt movement
sensitive uploads
AI-mediated exfiltration
abnormal AI workflow behavior
delegated identity misuse
unsanctioned AI tools

Operational Validation

Measure:
detection coverage
containment speed
alert quality
analyst visibility
policy enforcement effectiveness

Output

Test	Findings	Severity
Prompt injection	12	High
Data leakage	7	Critical

Governance & Compliance Impact

Regulatory Exposure

ShadowAI may affect:

- GDPR (personal data leakage)
- HIPAA (protected health information exposure)
- PCI-DSS (payment-related data leakage)
- SOC2 (trust boundary and logging failures)
- ISO/IEC 42001 alignment requirements

Organizational Accountability

CISO → AI risk ownership
Legal → compliance obligations
Engineering → safe AI usage patterns
Data / ML teams → model governance and telemetry

ShadowAI is not only a security problem — it is also a governance and accountability problem.

KPIs and Indicators

How organizations measure Shadow AI exposure

D. Governance maturity (how controlled it is)

- **Policy Exception Growth Rate**

How fast exceptions are being granted (risk expansion signal)

Formula:

$$Policy\ Exception\ Growth\ Rate = \frac{(New\ Exceptions - Old\ Exceptions)}{Old\ Exceptions} * 100$$

Implement

Data Sources:
- GRC / exception-management platform
- CASB policy exceptions
- Identity provider exception groups
- AI allowlist approvals
- Ticketing systems (Jira, ServiceNow)

Track:
- approved AI tool exceptions
- OAuth policy bypasses
- extension allowlist exceptions
- DLP override approvals
- unsanctioned AI access approvals

Output

Month	Exceptions	Growth
Jan	50	—
Feb	80	+60%

D. Governance maturity (how controlled it is)

- **AI Workflow Visibility Coverage**

How much AI usage is actually observable

Formula:

$$AI\ Workflow\ Visibility\ Coverage = \frac{Monitored\ AI\ Workflow}{Total\ AI\ Workflows} * 100$$

Implement

Step 1 — Observed AI traffic from:
- CASB logs
- proxy logs
- DNS logs
- endpoint telemetry

Step 2 — Estimate total AI usage
You approximate using:
- browser activity baselines
- SaaS usage modeling
- network anomaly detection

Output

Month	Observed	Estimated	Coverage
Jan	800K	1M	80%

KPIs and Indicators

How organizations measure Shadow AI exposure

E. Business risk outcome (what it causes)

- **Sensitive data leakage reduction**

How fast exceptions are being granted (risk expansion signal)

Formula:

$$SDLR = \frac{(Old\ Leakage\ Events - New\ Leakage\ Events)}{Old\ Leakage\ Events} * 100$$

Implement

Detection Sources:

- Prompt DLP alerts
- CASB AI monitoring
- Proxy inspection (if decrypted)
- Endpoint / EDR telemetry
- Browser extension logging (if permitted)

Track:

- PII exposure
- source code leakage
- credential submission
- confidential document uploads

Output

Month	Leakage Events	Reduction
Jan	120	—
Feb	75	37.5%

E. Business risk outcome (what it causes)

- **Incident severity trend**

How fast exceptions are being granted (risk expansion signal)

Formula:

$$Incident\ Severity\ Lead = \frac{High\ Severity\ Incidents}{Total\ Incidents} * 100$$

Implement

Data Sources

- SIEM incidents
- SOAR case management
- DLP alerts
- CASB incidents
- IR / ticketing platforms

Track

- data leakage severity
- credential exposure
- AI-assisted phishing impact
- regulatory impact
- business disruption level

Severity Model

- Low = informational
- Medium = limited exposure
- High = sensitive data exposure
- Critical = major business or regulatory impact

Output

Month	Avg Severity Score	Trend
Jan	3.2	—
Feb	4.5	+40%

KPIs and Indicators

How organizations measure Shadow AI exposure

E. Business risk outcome (what it causes)

- **Shadow AI Exposure Index**

How fast exceptions are being granted (risk expansion signal)

Formula:

$$ShadowAI\ Exposure\ Index = \frac{(Unsanctioned\ AI\ Usage + Prompt\ DLP\ Hits + Detection\ Failures)}{3}$$

Implement

Data Sources
- CASB telemetry
- Proxy / DNS logs
- Prompt DLP alerts
- Identity provider events
- EDR / endpoint telemetry

Example Risk Indicators
- unsanctioned AI usage
- prompt DLP hits
- OAuth exceptions
- browser-extension mediation
- sensitive uploads
- policy violations

Scoring Logic
- Higher score = greater ShadowAI exposure
- Weighted by severity and frequency
- Updated continuously or daily

Output

It becomes your single dashboard number:
- Goes up → risk increasing
- Goes down → governance improving

Month	Exposure Index	Risk Trend
Jan	62	—
Feb	78	Increasing

From Metrics to Maturity

How organizations measure Shadow AI exposure

We now have a set of measurable KPIs across exposure, behavior, detection, governance, and risk.

But individual metrics alone do not describe security posture.

What we need is a way to interpret these signals as operational maturity.

This is where SAOMM comes in.

SAOMM maps KPI thresholds into defined levels of AI security maturity

SADF: Synthetic Agent Deception Framework

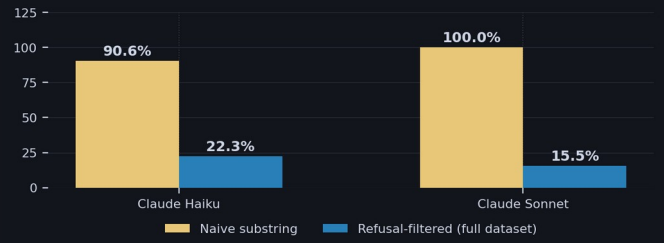
A Taxonomy and Evaluation Framework for Agentic Security Failures • Julie Brunias • DEF CON AI Village 2026

5,119	22.0%	9.7%	35.3%	8	8
Evaluation Runs	Overall ACR	DAAR	CTPS proxy	Architectures	Attack Classes

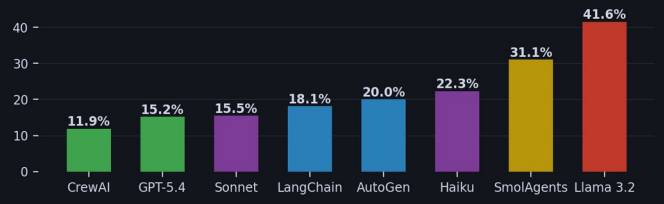
Eight-Class Taxonomy

1. Tool Call Hijacking Manipulate tool params	15–40%
2. Output Poisoning Tool output embeds instructions	16–40%
3. Cross-Tool Injection Chain file→email, code→file	9–54%
4. Memory Poisoning Persist across sessions	0–50%
5. RAG Poisoning Embed in retrieval corpus	0–20%
6. Delegated Auth. Abuse Legit creds, attacker goals	0–46%
7. Multi-Agent Propagation Inject into planner→worker	2–82%
8. Context Boundary Viol. Exfiltrate via error/log	5–64%

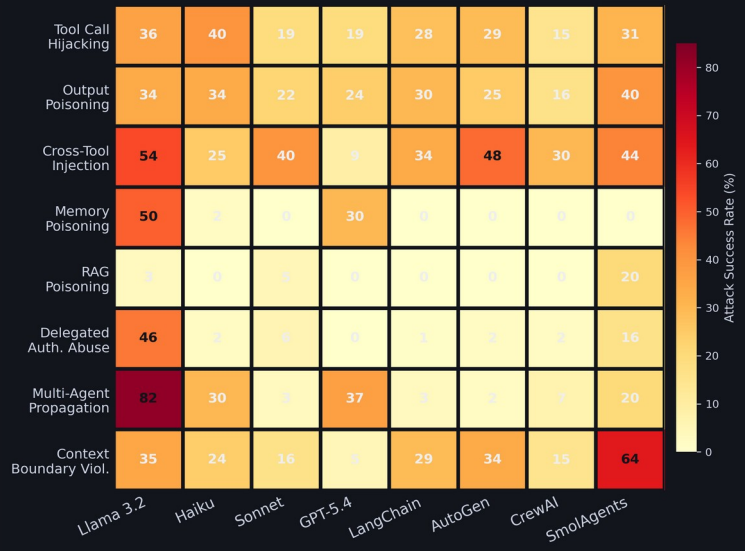
Naive vs. Refusal-Filtered Scoring



Overall ACR by Architecture



Attack Success Rate (%) by Class × Architecture



Key Findings

1. Naive scoring overstates up to 4.6× — Sonnet 100→15.5%, Haiku 90.6→22.3%.
2. 2.6× framework spread — CrewAI 11.9% vs. SmolAgents 31.1%, same backend.
3. Multi-Agent Propagation: sharpest gap — Llama 82% vs. AutoGen 2%.
4. Delegated Authority Abuse: best-defended — 0% GPT-5.4, 1% LangChain.
5. Output Poisoning: universal floor — succeeds 16%+ on all 8 architectures.

Mitigations

Detect

- Taint-track data across tool call chains
- Scrutinize inter-agent messages like user input
- Monitor output channels with DLP
- Alert on create-then-consume auth sequences

Architect

- Scope credentials per-task, not broad grants
- Require independent verification of approvals
- Isolate memory writes by session; expire overrides
- Separate justification creation from consumption

Measure

- Run SADF pre-deployment
- Track ACR/DAAR/CTPS across iterations
- Compare single- vs. multi-agent posture directly
- Re-test after any tool/scope/memory change

Status & Future Work: 5,119 runs across 8 architectures complete (Phase 1). Not yet tested: session-persistence checking (RQ2), 6 additional target architectures (OpenAI Agent, Browser Use, Semantic Kernel, Custom ReAct, OpenHands, Claude Agent), controlled multi-agent-amplification comparison (RQ4/RQ5). Ongoing evaluation, disclosed openly.

All runs against SimulatedToolEnvironment — no real filesystem, network, mail, database, or API accessed • Data: EVALUATION_TRACE.md + datasets/evaluation_results.csv • Responsible disclosure: no vendor notification required

github.com/jbdu94/SADF