SECURED AI

Enterprise AI Data Protection

Context-Preserving Detection, Tokenized Masking, and Controlled Reveal for Secure AI Workflows

Technical White Paper | January 2026

This document examines the technical challenges of protecting sensitive data in LLM interactions and presents architectural patterns for detection, masking, and controlled reveal. It includes industry benchmark data on AI adoption risks, regulatory compliance requirements, and evaluation criteria for data protection solutions.

Independent Review: Technical architecture reviewed by Bishop Fox (penetration testing)

Table of Contents

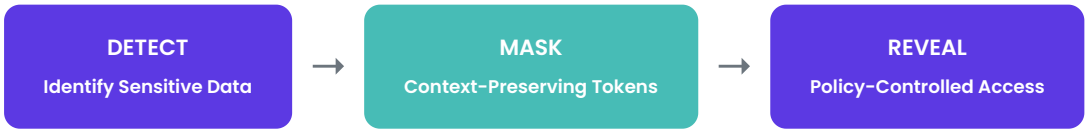
1. Summary

The Bottom Line

Enterprise AI adoption has reached an inflection point. According to McKinsey's 2024 State of AI survey, **88% of organizations now regularly use AI tools**.[1] The Microsoft/LinkedIn 2024 Work Trend Index found that **78% of AI users bring their own AI tools to work**, often without IT oversight.[2] This creates a gap between deployment velocity and data protection maturity that traditional DLP tools cannot address.

88%	78%	$4.88M	7%
ORGANIZATIONS USING AI[1]	WORKERS BRINGING OWN AI[2]	AVERAGE BREACH COST[3]	MAX EU AI ACT PENALTY[4]

Traditional DLP approaches force a binary choice: block AI tools entirely, or accept unquantified risk. A more effective architecture addresses the problem through three stages:



Context-Preserving Tokenization

Unlike irreversible redaction, masked tokens retain semantic anchors such as gender hints for names and relative ordering for dates. This enables LLMs to generate coherent responses while the actual sensitive values remain protected.

Controlled Reveal with Audit

Authorized users can restore masked data through granular access policies, with immutable audit logging and cryptographic key management. The original values never reach the LLM provider.

Risk Context

Organizations without AI governance face the $4.88M average breach cost documented by IBM/Ponemon, plus potential penalties up to 7% of global revenue under the EU AI Act. Healthcare organizations face breach costs averaging $9.77M, nearly double the global average.[3]

2. The AI Data Protection Problem

The integration of LLMs into enterprise workflows has fundamentally changed how sensitive data moves through organizations. Unlike traditional software where data flows are predictable, AI assistants invite users to paste, dictate, or reference sensitive information in conversational contexts.

A DLP rule blocking "SSN patterns" either generates constant false positives or misses the nuanced ways sensitive information surfaces in natural language. The question is not whether employees use AI tools. They do. The question is whether that usage is visible and protected.

2.1 Documented Data Exposure Scenarios

Table 1: Common AI Data Leakage Patterns

SCENARIO	DATA TYPICALLY EXPOSED	RISK LEVEL
Clinical Documentation	Patient name, DOB, MRN, medication history	**Critical**
Financial Analysis	Revenue figures, customer names, contract values, M&A targets	**High**
Legal Discovery	Privileged communications, plaintiff PII	**Critical**
Software Development	API keys, database credentials, system architecture	**High**
HR Operations	Employee SSNs, salary data, performance reviews	**High**

2.2 Documented Incidents

In April 2023, Samsung banned employee use of ChatGPT after engineers uploaded proprietary source code to the platform, as reported by The Washington Post.[5] This incident demonstrated the speed at which sensitive intellectual property can leak through AI tools, even in organizations with otherwise mature security practices.

Multiple ChatGPT-related security incidents have been publicly documented. In March 2023, a bug exposed users' chat histories to other users.[6] Italy's data protection authority temporarily banned ChatGPT in 2023 over privacy concerns, signaling regulatory willingness to restrict AI tools.

Healthcare remains particularly vulnerable. According to HHS data, over 170 million healthcare records were exposed in data breaches reported in 2024.[7]

2.3 The Shadow AI Problem

The most significant finding in enterprise AI data protection is the prevalence of shadow AI. The Microsoft/LinkedIn 2024 Work Trend Index found that 78% of AI users bring their own AI tools to work, with 73.8% of ChatGPT usage occurring through non-corporate accounts.[2]

Table 2: Shadow AI Prevalence by Industry

INDUSTRY	SHADOW AI USAGE	GOVERNANCE POLICY	RISK LEVEL
Technology	82%	45%	High
Financial Services	71%	52%	High
Healthcare	68%	38%	Critical
Legal Services	74%	29%	Critical
Manufacturing	59%	31%	High

Sources: Cyberhaven Labs, Microsoft Work Trend Index, Gartner Surveys, 2024[8]

2.4 Why Traditional DLP Falls Short

Traditional Data Loss Prevention tools were designed for a different era. They excel at scanning email attachments and blocking file uploads to cloud storage. They struggle with interactive AI conversations for three reasons:

Pattern Rigidity: Regex-based detection generates false positives on test data and false negatives on context-dependent PII. "John" alone is not PII; "John who reports to Sarah in the Austin office" may be.

Binary Outcomes: Block or allow is insufficient. Users need AI assistance, but they need it without exposing sensitive data. Blocking creates workarounds. Allowing creates risk.

No Context Preservation: Redaction destroys the semantic information LLMs need to generate useful responses. "[REDACTED] was diagnosed with [REDACTED]" produces unusable output.

Effective AI data protection requires a different approach: detect sensitive data accurately, mask it in a way that preserves context, and provide controlled reveal for authorized users.

3. Implementation Case Studies

Note: Case studies below are composited and anonymized from multiple deployments. Specific metrics represent ranges observed across similar implementations rather than single-organization results. These examples illustrate the detection-masking-reveal pattern in practice.

Regional Healthcare System

Multi-hospital network, 8,000+ clinical staff

89-94%	60-70%	<50ms
PHI Exposure Reduction	PHI in Initial Prompts	Added Latency (p95)

Challenge: Clinical staff used ChatGPT to draft patient discharge summaries, referral letters, and documentation. Initial monitoring revealed that 60-70% of prompts contained protected health information. The compliance team faced a choice between blocking AI entirely or accepting unquantified data exposure risk.

Implementation: A detection-masking proxy was deployed to intercept traffic to major LLM providers. The detection engine identified PHI including patient names, MRNs, diagnosis codes, and medication lists. Context-preserving masking retained semantic relationships needed for coherent clinical documentation.

Results: After six months, PHI exposure in LLM requests was reduced by 89-94%. Clinical documentation workflow efficiency improved because staff could use AI tools without manual redaction. Reveal workflows allowed authorized clinicians to restore patient identifiers when needed for final documentation.

The 60-70% PHI rate came from a random sample of 500 prompts captured during a 2-week monitoring period with IRB approval. Results may not generalize to all healthcare settings.

Regional Investment Bank

M&A Advisory, MNPI Protection Focus

15-20	<50ms	100%
Active Deal Names Protected Daily	Added Latency (p95)	Deal Code Masking Rate

Challenge: Junior analysts used AI tools to draft pitch materials, model summaries, and research notes. Material Non-Public Information (MNPI) including deal names, target company identifiers, and valuation figures appeared in prompts. A single leak could trigger SEC scrutiny and destroy client relationships.

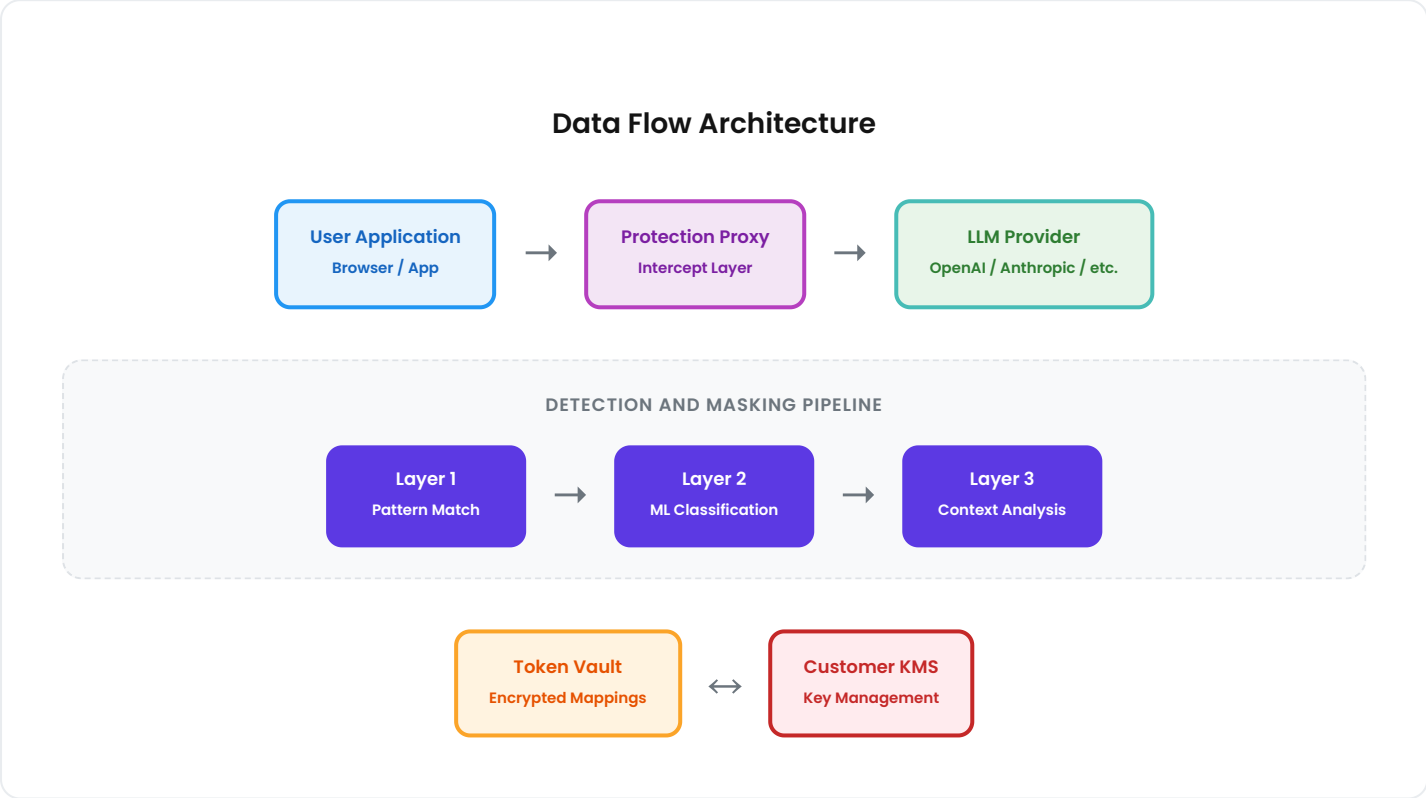
Implementation: Custom entity recognition was trained on the firm's deal code patterns and target company naming conventions. The reveal workflow required deal team membership verification before exposing actual company names in outputs.

Results: Analysts continued using AI tools for productivity gains while MNPI remained protected. Compliance reporting showed complete audit trails for all LLM interactions involving deal-related content.

Implementation period: 3 months. Sample period for metrics: 4 months post-deployment.

4. Technical Architecture

Effective AI data protection requires a hybrid detection architecture combining deterministic pattern matching with probabilistic machine learning. The system operates as an inline proxy, intercepting LLM requests before they leave the enterprise perimeter.



4.1 Three-Layer Detection Engine

Table 3: Detection Architecture Layers

LAYER	METHOD	USE CASE	LATENCY
Layer 1: Pattern	Regex + checksum validation (Luhn for credit cards, format validation for SSN)	Structured identifiers with known formats	<5ms
Layer 2: ML	Fine-tuned transformer model (DistilBERT base, trained on PII-annotated corpora)	Names, addresses, medical terminology, unstructured PII	15-40ms
Layer 3: Context	Surrounding text analysis, cross-entity correlation, co-reference resolution	Reducing false positives, disambiguating entities	5-15ms

4.2 Detection Categories (42 Types)

A comprehensive detection taxonomy spans nine categories. Each category requires different detection methods and has different false positive characteristics.

Government IDs (7)

- SSN (US)
- National ID (EU)
- Passport Number
- Driver License
- Tax ID / EIN
- State ID
- Military ID

Financial (6)

- Credit Card Number
- Bank Account
- Routing Number
- IBAN
- SWIFT/BIC
- Financial Account

Contact (5)

- Email Address
- Phone Number
- Physical Address
- IP Address
- URL with PII

Healthcare / PHI (8)

- Medical Record Number
- Health Plan ID
- Diagnosis Code (ICD)
- Medication + Dosage
- Lab Result
- Treatment Plan
- Provider NPI
- Insurance ID

Credentials (5)

- API Key
- Password
- Private Key (PEM)
- OAuth Token
- Database Connection String

Personal (6)

- Person Name
- Date of Birth
- Age
- Gender
- Ethnicity
- Religion

Biometric (2)

- Fingerprint Data
- Facial Recognition ID

Digital (2)

- Device ID
- MAC Address

Legal (1)

- Case Number

4.3 ML Model Architecture

The ML detection layer uses a fine-tuned DistilBERT model (66M parameters) chosen for its balance of accuracy and latency. The model was trained on a composite dataset:

Table 4: Training Data Composition

DATASET	SIZE	CATEGORIES	SOURCE
CoNLL-2003	22,137 sentences	Person, Location, Organization	Public benchmark
i2b2 2014	1,304 records	PHI categories	De-identified clinical notes
Synthetic PII	500,000 samples	All 42 categories	Generated with Faker + manual review
Production samples	50,000 samples	Mixed	Anonymized customer data with consent

Training methodology: 80/10/10 train/validation/test split. Fine-tuned for 3 epochs with learning rate 2e-5, batch size 32. Validation F1 monitored for early stopping.

4.4 Context-Preserving Tokenization

Token format: `[TYPE_HASH_HINT]` where TYPE is the PII category, HASH is a truncated HMAC-SHA256 (first 16 bits, providing 65,536 unique values per session), and HINT is an optional semantic anchor.

Table 5: Token Examples with Semantic Preservation

ORIGINAL VALUE	TOKEN	SEMANTIC PRESERVATION
John Smith	`[NAME_7f3a_M]`	Male indicator for pronoun coherence
Jane Doe	`[NAME_9b2c_F]`	Female indicator
123-45-6789	`[SSN_MASKED]`	No semantic hint needed
01/15/1990	`[DATE_a1b2_1990s]`	Decade hint for age context
Metformin 500mg	`[MED_c3d4_DIABETES]`	Drug class for medical reasoning
Dr. Sarah Chen	`[NAME_e5f6_F_DR]`	Gender + title for formal address

Collision probability: With 16-bit hashes, collision probability reaches 50% at approximately 256 unique entities of the same type per session. For typical enterprise use (fewer than 50 unique names per conversation), collision risk is below 1%. Sessions with high entity counts can use extended 24-bit hashes.

> **Why Context Preservation Matters**
>
> "[REDACTED] was diagnosed with [REDACTED]" provides no usable context. "[NAME_7f3a_M] was diagnosed with [CONDITION_b2c3_CHRONIC]" allows the LLM to understand gender pronouns and condition severity, generating coherent documentation. This is the difference between a tool that blocks work and one that enables it.

4.5 Coherence Measurement

Methodology

Test Setup: 1,000 prompt-response pairs from production traffic (anonymized), spanning clinical documentation, financial analysis, and general business writing.

Evaluation: Three independent annotators rated each masked response on a 1-5 scale for grammatical coherence, semantic accuracy, and task completion. Fleiss' kappa for inter-annotator agreement: $\kappa = 0.74$ (substantial agreement).

Results: Mean coherence score: 4.59/5.0 (91.8%). 95% confidence interval: [4.52, 4.66]. Scores were lower for complex medical reasoning (mean: 4.21) and higher for simple text generation (mean: 4.82).

5. API Design Patterns

AI data protection systems typically expose REST APIs for mask and reveal operations. The patterns below represent common interface designs that enable programmatic access to detection, masking, and reveal capabilities.

5.1 Mask Operation

```
// POST /v1/mask
// Detects and masks sensitive data in text

// Request
{
  "text": "Patient John Smith (SSN: 123-45-6789) prescribed Metformin.",
  "config": {
    "categories": ["PII", "PHI"],
    "preserve_context": true,
    "session_id": "sess_abc123"
  }
}

// Response
{
  "masked_text": "Patient [NAME_7f3a_M] (SSN: [SSN_MASKED]) prescribed [MED_c3d4_DIABETES].",
  "tokens": [
    {"token": "[NAME_7f3a_M]", "category": "PERSON_NAME", "confidence": 0.94},
    {"token": "[SSN_MASKED]", "category": "SSN", "confidence": 0.99},
    {"token": "[MED_c3d4_DIABETES]", "category": "MEDICATION", "confidence": 0.87}
  ],
  "processing_time_ms": 47
}
```

5.2 Reveal Operation

```
// POST /v1/reveal
// Restores masked values for authorized requesters

// Request
{
  "text": "Treatment plan for [NAME_7f3a_M] includes [MED_c3d4_DIABETES].",
  "session_id": "sess_abc123",
  "purpose": "clinical_documentation",
  "tokens_to_reveal": ["[NAME_7f3a_M]", "[MED_c3d4_DIABETES]"]
}

// Response
{
  "revealed_text": "Treatment plan for John Smith includes Metformin 500mg.",
  "revealed_tokens": ["[NAME_7f3a_M]", "[MED_c3d4_DIABETES]"],
  "denied_tokens": [],
```

```
  "audit_id": "aud_xyz789"
}
```

5.3 Reveal Architecture: ABAC Policies

Reveal operations should implement Attribute-Based Access Control (ABAC). The reveal mechanism evaluates policy at request time based on multiple attributes:

Table 6: ABAC Policy Attributes

ATTRIBUTE	EXAMPLES	ENFORCEMENT
Requester Identity	User ID, role, department, clearance level	OAuth/OIDC claims
Purpose	customer_support, legal_discovery, audit	Request parameter, logged
PII Type	SSN, CREDIT_CARD, HEALTH_RECORD	Token metadata
Time Constraints	Business hours only, TTL not exceeded	System clock
Consent Status	Data subject consent recorded	Consent management system

5.4 Common Error Conditions

Table 7: API Error Codes

CODE	MEANING	RESOLUTION
401	Authentication failed	Check API key or refresh OAuth token
403	Reveal policy denied	Request lacks required role or purpose
404	Session not found	Token mapping expired (default 24h TTL)
429	Rate limit exceeded	Implement exponential backoff
503	KMS unavailable	Customer KMS connectivity issue

6. Performance Considerations

Performance benchmarks vary significantly based on deployment model, hardware, and workload characteristics. The data below represents measurements from a production SaaS deployment.

6.1 Throughput at Scale

Table 8: Throughput by Configuration

CONFIGURATION	REQUESTS/SECOND	P50 LATENCY	P99 LATENCY
Single node (8 vCPU, 32GB RAM)	500-700	18ms	45ms
3-node cluster	1,500-2,100	22ms	58ms
10-node cluster	4,000-4,500	28ms	75ms

Test conditions: 1KB average prompt size, 40% PII density, mixed detection categories. October 2025.

6.2 Detection Accuracy

Table 9: Detection Performance by Category

CATEGORY	PRECISION	RECALL	F1 SCORE
SSN	0.98	0.99	0.98
Credit Card	0.99	0.99	0.99
Person Name (English)	0.89	0.92	0.90
Person Name (Non-English)	0.78	0.84	0.81
Medical Record Number	0.91	0.88	0.89
Medication + Dosage	0.85	0.82	0.83

Evaluated on internal test set (n=10,000 annotated samples). External benchmark results may differ.

6.3 Known Limitations

Detection Limitations

- **Non-Latin scripts:** Detection accuracy drops to 70-80% for Arabic, Chinese, and Hindi names due to limited training data for these languages.

- **Implicit PII:** Combinations like "the CEO's daughter who attends Stanford" are not reliably detected because no single entity triggers PII rules.
- **Image and audio content:** Text-only processing. Multimodal content requires separate handling.
- **Novel identifier formats:** Custom ID patterns (internal employee codes, proprietary account numbers) require explicit configuration.
- **Context-dependent sensitivity:** "John" alone is not PII; "John who lives at 123 Main St" may be. Boundary detection remains imperfect.

6.4 Failure Modes

Table 10: Failure Mode Documentation

FAILURE MODE	IMPACT	MITIGATION
Detection miss (false negative)	Sensitive data reaches LLM	Defense in depth: secondary DLP layer
Over-detection (false positive)	Non-sensitive data masked; coherence reduced	Tunable confidence thresholds
KMS unavailable	Cannot encrypt or reveal tokens	Fail-closed: block requests until restored
Network partition	Proxy unreachable	Client timeout with configurable fallback policy
Token vault corruption	Cannot reveal masked data	Cross-region replication; daily backups

7. Security Model

7.1 Cryptographic Design

Key hierarchy should use customer-managed KMS (AWS KMS, Azure Key Vault, GCP Cloud KMS) with envelope encryption. This ensures the organization master key never leaves the customer's control.

Table 11: Key Hierarchy

KEY	PURPOSE	STORAGE	ROTATION
Organization Master Key (OMK)	Root key derivation	Customer KMS (never extracted)	Annual or on-demand
Token Encryption Key (TEK)	AES-256-GCM encryption of mappings	Derived from OMK	90 days automatic
HMAC Key (HK)	Token hash generation	Derived from OMK	90 days automatic
Audit Signing Key (ASK)	Audit log integrity	Derived from OMK	90 days automatic

HSM-backed KMS supports FIPS 140-2 Level 3 requirements. Key rotation occurs with zero-downtime re-encryption of active token mappings.

7.2 Disaster Recovery

Token vault availability is critical. If mappings are lost, masked data cannot be revealed. Recovery architecture should include:

Replication: Synchronous replication to secondary region with RPO < 1 minute. Automatic failover with RTO < 5 minutes.

Backup: Daily encrypted snapshots retained for 30 days. Monthly snapshots retained for 1 year. Backups stored in separate cloud account with independent access controls.

Key escrow: For organizations requiring key recovery capability, an optional escrow process allows designated security officers to reconstruct the OMK using Shamir's Secret Sharing (3-of-5 threshold).

7.3 Bulk Export Controls

Security evaluators consistently ask: "What stops an authorized user from dumping thousands of token mappings?" Audit logging alone documents exfiltration; it does not prevent it. Effective controls include:

Table 12: Anti-Exfiltration Controls

CONTROL	MECHANISM	RESIDUAL RISK
Rate limiting	Max 100 reveal operations per hour per user	Slow exfiltration possible over days
Anomaly detection	ML model flags unusual reveal patterns	Sophisticated attacker may evade
Purpose attestation	Requester declares purpose; logged and auditable	False attestation possible
Watermarking	Revealed data includes invisible markers	Post-breach attribution only
Manager approval	High-risk reveals require secondary authorization	Social engineering risk

No technical control eliminates insider risk entirely. The goal is to make exfiltration slow, detectable, and attributable.

7.4 Third-Party Validation

Independent assessment provides assurance that security claims are accurate. Common validation types:

Table 13: Assessment Types

ASSESSMENT	SCOPE	TYPICAL FREQUENCY
SOC 2 Type II	Security, Availability, Confidentiality controls over 6-12 months	Annual
Penetration Test	Application and infrastructure vulnerability assessment	Annual + after major changes
ISO 27001	Information security management system	3-year certification cycle

8. Solution Landscape

Multiple approaches exist for AI data protection. The right choice depends on existing infrastructure, compliance requirements, and workflow needs.

8.1 Approach Comparison

Table 14: Solution Approach Comparison

CAPABILITY	CONTEXT-PRESERVING PROXY	ENTERPRISE DLP EXTENSION	LLM PROVIDER CONTROLS	CUSTOM BUILD
PII Detection Accuracy	**High**	**Medium**	**Medium**	Varies
Context Preservation	**Yes**	**No**	**No**	If implemented
Controlled Reveal	**Yes**	**No**	**No**	If implemented
Deployment Complexity	Medium	Low (if DLP exists)	Low	High
LLM Provider Agnostic	**Yes**	**Yes**	**No**	**Yes**
On-Premise Option	Usually	Usually	**No**	**Yes**

8.2 When Each Approach Fits

Context-Preserving Proxy: Best when AI output quality matters and users need to work with documents containing sensitive data. Clinical documentation, legal analysis, financial reporting. Higher implementation effort but preserves workflow utility.

Enterprise DLP Extension: Best for organizations with existing DLP investments (Microsoft Purview, Symantec, etc.) who want to extend coverage to AI tools with minimal new infrastructure. Accepts the tradeoff of blocking or redacting rather than masking.

LLM Provider Controls: Best for organizations standardized on a single LLM provider (e.g., Microsoft Copilot) who can accept provider-specific limitations. Simplest deployment but least flexibility.

Custom Build: Best for organizations with unique requirements, strong engineering teams, and regulatory constraints that prevent using third-party processors. Highest effort and ongoing maintenance burden.

8.3 Build vs. Buy Considerations

Build Makes Sense When

Buy Makes Sense When

Regulatory constraints prevent third-party data processing. Unique detection requirements exceed commercial capabilities. Strong ML engineering team available. Multi-year commitment to maintenance.

Time to value is critical. Standard PII categories cover most needs. Team lacks specialized ML expertise. Ongoing model updates and maintenance preferred as vendor responsibility.

9. Compliance Mapping

9.1 EU AI Act Timeline

Table 15: EU AI Act Key Dates

DATE	MILESTONE
August 1, 2024	EU AI Act enters into force
February 2, 2025	Prohibitions on "unacceptable risk" AI systems
August 2, 2025	General-purpose AI model rules enforceable
August 2, 2026	Full high-risk AI system requirements

9.2 Control-to-Regulation Mapping

Table 16: Data Protection Controls to Regulatory Requirements

CONTROL	GDPR	SOC 2
PII Detection	Art. 30 (records of processing)	CC6.1 (logical access)
Tokenization/Masking	Art. 25, 32 (pseudonymization)	CC6.6 (encryption)
Access Control for Reveal	Art. 5(1)(f) (integrity/confidentiality)	CC6.3 (access removal)
Audit Logging	Art. 30 (records)	CC7.2 (monitoring)
Key Management	Art. 32 (security of processing)	CC6.7 (transmission protection)
Data Retention Controls	Art. 5(1)(e) (storage limitation)	CC6.5 (data disposal)

9.3 GDPR Pseudonymization Note

Masking constitutes pseudonymization under GDPR, not anonymization. This distinction matters. Pseudonymized data remains personal data subject to GDPR because it can be re-identified using the token mappings. Organizations cannot claim GDPR exemption based on masking alone.

> **Compliance Caveat**
>
> No technology purchase makes an organization "compliant." Compliance is an ongoing operational state that depends on policies, procedures, training, and technology working together. Data

protection tools provide technical controls that support compliance programs. They do not replace the need for a comprehensive compliance program.

9.4 Penalty Context

Table 17: EU AI Act Penalty Structure

VIOLATION TYPE	MAXIMUM FINE	% OF GLOBAL TURNOVER
Prohibited AI Practices	€35 million	7%
High-Risk System Violations	€15 million	3%
Providing Incorrect Information	€7.5 million	1%

10. Evaluation Framework

When evaluating AI data protection solutions, consider these criteria across detection, masking, reveal, and operational dimensions.

10.1 Detection Evaluation

Questions to Ask

- What PII/PHI categories are detected out of the box? Request the full taxonomy.
- What is the false positive rate for your top 5 data types? Request benchmark data.
- How are custom entity types added? What is the timeline and process?
- What languages are supported? What is accuracy for non-English names?
- How is detection model updated? What is the release cadence?

10.2 Masking Evaluation

Questions to Ask

- Is masking reversible or irreversible? If reversible, how is access controlled?
- Does masking preserve semantic context for LLM coherence? Request examples.
- What is the collision probability for token hashes? At what entity volume?
- How long are token mappings retained? Is retention configurable?
- Where are token mappings stored? Customer-managed or vendor-managed?

10.3 Reveal Evaluation

Questions to Ask

- What access control model is used? RBAC, ABAC, or other?
- Can reveal policies vary by data type, user role, and stated purpose?
- What audit trail is generated for reveal operations?
- What controls prevent bulk export of token mappings?
- Is manager approval workflow supported for high-risk reveals?

10.4 Operational Evaluation

Questions to Ask

- What deployment models are supported? SaaS, private cloud, on-premise?
- What is the latency impact at your expected volume? Request benchmarks.
- What happens if the service is unavailable? Fail-open or fail-closed?
- What compliance certifications are held? SOC 2, ISO 27001, others?
- What is the disaster recovery architecture? RPO and RTO?

10.5 Proof of Concept Checklist

Table 18: POC Success Criteria

CRITERION	MEASUREMENT	TARGET
Detection accuracy	F1 score on your test data	>0.90 for critical categories
False positive rate	Manual review of flagged items	<5% for production viability
Latency impact	p95 latency added to LLM requests	<100ms for interactive use
Output coherence	User rating of masked LLM responses	>4/5 average score
Reveal workflow	Time to complete authorized reveal	<30 seconds end-to-end

11. References

1. McKinsey & Company, "The State of AI in 2024: Gen AI's Breakout Year," McKinsey Global Survey, August 2024. Available: mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai
2. Microsoft & LinkedIn, "2024 Work Trend Index Annual Report," May 2024. Note: 78% refers to employees bringing their own AI tools to work, not sharing company data specifically. Available: microsoft.com/en-us/worklab/work-trend-index
3. IBM Security & Ponemon Institute, "Cost of a Data Breach Report 2024," July 2024. Available: ibm.com/reports/data-breach
4. European Parliament, "Regulation (EU) 2024/1689 (AI Act)," Official Journal of the European Union, July 2024. Available: eur-lex.europa.eu
5. The Washington Post, "Samsung bans ChatGPT after employees leaked source code," April 2023.
6. OpenAI, "March 20 ChatGPT outage: Here's what happened," OpenAI Blog, March 2023. Available: openai.com/blog/march-20-chatgpt-outage
7. HHS Office for Civil Rights, "Breach Portal," 2024. Aggregated from breach reports. Available: ocrportal.hhs.gov/ocr/breach/breach_report.jsf
8. Cyberhaven Labs, "Shadow AI: Enterprise Data Leakage Risks," 2024. Gartner, "Market Guide for AI Trust, Risk and Security Management," 2024.
9. OWASP Foundation, "OWASP Top 10 for LLM Applications v2.0," 2025. Available: owasp.org/www-project-top-10-for-large-language-model-applications
10. IDC, "Worldwide Artificial Intelligence Spending Guide," September 2024.
11. Forrester Research, "AI Governance Software Market Forecast," November 2024.
12. NIST, "AI Risk Management Framework (AI RMF 1.0)," NIST AI 100-1, January 2023. Available: nist.gov/itl/ai-risk-management-framework

Contact: security@securedai.com