OPSWAT.

SOLUTION BRIEF

METADEFENDER CORE™

Securing AI Data Pipelines and LLM-Powered Applications

Protecting the World's Critical Infrastructure

AI Adoption at Scale and the Expanding Attack Surface

Enterprise AI adoption is accelerating rapidly across industries. With organizations exploring or deploying AI in production, security and compliance become the top barriers to adoption. As AI systems ingest increasing volumes of external data including documents, APIs, files, and third-party content, the risk of threat amplification and data compromise grows significantly.

LLM-powered applications are particularly vulnerable, as they process unstructured, often unverified content at high speed and scale. Without proper controls, these systems may inadvertently consume malware, leak sensitive data, or become targets for model manipulation and data poisoning.

Organizations must secure AI data intake with the same rigor applied to critical infrastructure and zero-trust architecture.

Data Intake Is the New AI Attack

AI systems consume content from diverse and dynamic sources:

- User file uploads (e.g., web pages, PDFs, image)
- Third-party APIs and data repositories
- Internal unstructured content (emails, reports, wikis)
- Fine-tuning datasets and real-time prompts

Each of these intake points presents a potential attack surface. Adversaries may exploit them to:

- Inject malware via embedded macros or scripts
- Tamper with training data to trigger model poisoning
- Embed sensitive information to trigger data leakage in outputs
- Deliver malformed files to exploit backend AI infrastructure

Sensitive data are continuously processed by LLM services including uploaded documents, user conversations, and content stored in persistent memory. Commonly targeted data types include:

- Personal data: Names, email addresses, phone numbers, and Social Security numbers (SSNs)
- Financial information: Bank account details, credit card numbers, and payment data
- Health information: Medical records and other protected health information (PHI)
- Business-sensitive content: Trade secrets, strategic plans, internal financial reports
- Authentication credentials: Passwords, API keys, access tokens, and session identifiers
- Uploaded files: Confidential documents, government records, and proprietary research materials

High-Risk LLM Applications

As LLM capabilities expand, certain application patterns are emerging as especially vulnerable to abuse. These include:

1 Enterprise Chatbots with File Uploads

Enterprise LLM chatbots often accept document uploads to answer questions or summarize content. Attackers can submit weaponized PDFs or Office files to deliver malware or extract sensitive system behavior through prompt injection.

2 RAG (Retrieval-Augmented)

RAG architecture combines LLMs with external document retrieval to produce context-aware responses. Without validating and sanitizing indexed files, these systems are prone to poisoning, hallucinations, or model corruption due to malicious or manipulated content.

3 LLM-based Agents and Multi-Step Planners

LLM agents that make autonomous decisions based on file inputs, such as reading instructions or generating code, can be manipulated through embedded commands, malformed inputs, or weaponized dependencies. The complexity of agent pipelines makes real-time file security essential.

Each of these applications relies on external data trust, which must be validated and enforced with technical controls at the ingestion layer.

Key Challenges in Securing AI/ML Training Data

Securing AI/ML training data workloads presents a unique combination of security, privacy, and compliance challenges:

Data Complexity

AI systems process unstructured, variable, and often opaque file formats.

- For example, PDFs with embedded images or proprietary CAD formats require specialized parsing that traditional security tools can't handle.

Content Obfuscation

Embedded invisible text, hidden prompts, or off-screen content that is ignored by humans but parsed by AI systems.

- Attackers can embed white-on-white text or hide prompts in image metadata to manipulate AI model behavior.

Volume and Velocity

High-speed, automated data intake rules out manual inspection.

- Organizations processing millions of training files daily cannot manually verify each input without grinding operations to a halt.

Supply Chain Exposure

File inputs may originate from unknown or dynamically generated sources.

- Third-party data brokers or web scraping operations may unknowingly introduce poisoned datasets from compromised sources.

Compliance Drivers

The EU AI Act, Cyber Resilience Act, and others demand input validation, transparency, and continuous risk monitoring including training data provenance, output predictability, and human oversight.

- High-risk AI systems must demonstrate complete training data lineage with penalties up to Euro 35M or 7% of global revenue.

	OWASP Top 10 for LLMs
MITRE	ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems)
databricks	Vendor AI Security Frameworks
	Vendor AI Security Frameworks
ISO	AI Management Systems AI Functional Safety
	UN Resolution on AI Systems
G7	Declaration on the usage of AI

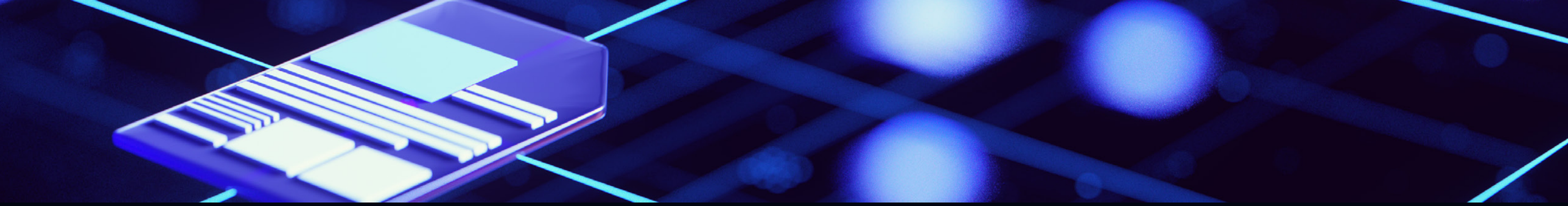
Global Compliance Drivers for AI Security

AI regulation is accelerating worldwide, imposing strict requirements for secure data handling, validated inputs, and continuous risk management. Leading frameworks such the EU AI Act and Cyber Resilience Act mandate that high-risk AI systems:

- Are protected against known vulnerabilities
- Validate all inputs and block malicious behavior
- Provide audit trails and enable post-deployment monitoring
- Embed data privacy, transparency, and governance by design

Protect AI/ML Model & Training Data with MetaDefender Core

MetaDefender Core delivers advanced threat detection and prevention for AI systems by integrating at key data intake points to ensure every file is clean, support compliance, and verify file integrity before use in AI workflows.



Enforce Zero-Trust for AI Data Ingestion

MetaDefender Core delivers multi-layered file security at the point of ingestion, ensuring every file is inspected, sanitized, and policy-enforced before it reaches training, inference, or indexing workflows. By treating all input as untrusted, it helps organizations apply Zero Trust principles to AI pipelines.

- File type verification detects spoofed or malformed files powered with AI.
- Country of Origin supports supply chain and geographic enforcement policies.
- Deep CDR™ removes embedded threats from documents and regenerates new, safe-to-use files.
- MetaScan™ Multiscanning utilizes 30+ anti-malware engines, as well as heuristics and machine learning, to identify over 99% of threats.
- Adaptive Sandbox emulates malicious behavior to detect evasive or zero-day malware.

Protecting Training Data and LLM Inputs

To mitigate poisoning, leakage, and integrity risks, MetaDefender Core secures fine-tuning datasets, real-time prompts, and indexed files across RAG, chatbot, and agent pipelines. This includes addressing emerging threats such as embedded prompts, document-layer obfuscation, and invisible text techniques where malicious content is hidden using formatting tricks to bypass human review while remaining fully accessible to AI systems during parsing and inference.

- File-based Vulnerability Assessment detects CVEs in model dependencies and tool chains.
- Proactive DLP™ scans for sensitive content (PII, PHI, financial data) and leverages Optical Character Recognition (OCR) technology to detect and redact hidden text within visual content.
- SBOM (Software Bill of Materials) generation helps identify components and track third-party software elements used in AI workloads.
- Fille Processing Workflow enables enforcement policies to ensure untrusted inputs are isolated or blocked.

Compliance with Data Privacy and Model Transparency Regulations

MetaDefender Core helps organizations align with global AI and data protection mandates, including the EU AI Act, Cyber Resilience Act, GDPR, HIPAA, and emerging frameworks in Asia-Pacific and North America. It enables secure input validation, full data processing traceability, and proactive risk mitigation for AI-generated outcomes.

- Enforces secure input validation and deep file inspection before ingestion.
- Provides complete audit trails, file hashes, and logging to support forensics and compliance reporting.
- Meets global privacy and transparency requirements, from EU to APAC to U.S. regulations.
- Supports proactive risk assessments across AI data supply chains to identify and remediate vulnerabilities early.

Deployment Flexibility for AI Environments

MetaDefender Core is a scalable, interoperable file security layer that is built to integrate with modern AI environments.

- Deploys anywhere whether cloud-native, on-premises, or air-gapped architectures.
- Integrates easily via REST API or ICAP-based integration with AI data ingestion flows, upload portals, AI training data pipelines, and LLM-powered applications.
- Scans at every stage from files in repositories to CI/CD pipelines used in AI models and chatbots development.
- Automates policy enforcement at the file-level based on contextual risk and operational requirements.

GET STARTED

Are you ready to put MetaDefender Core on the front lines of your cybersecurity strategy?

Talk to one of our experts today.

Scan the QR code or visit us at:
opswat.com/get-started
sales@opswat.com

For the last 20 years OPSWAT, a global leader in IT, OT, and ICS critical infrastructure cybersecurity, has continuously evolved an end-to-end solutions platform that gives public and private sector organizations and enterprises spanning Financial Services, Defense, Manufacturing, Energy, Aerospace, and Transportation Systems the critical advantage needed to protect their complex networks from cyberthreats.

Built on a "Trust no file. Trust no device.™" philosophy, OPSWAT solves customers' challenges like hardware scanning to secure the transfer of data, files, and device access with zero-trust solutions and patented technologies across every level of their infrastructure. OPSWAT is trusted globally by more than 1,800 organizations, governments, and institutions across critical infrastructure to help secure their devices, files, and networks from known and unknown threats, zero-day attacks, and malware, while ensuring compliance with industry and government-driven policies and regulations.

Discover how OPSWAT is protecting the world's critical infrastructure and securing our way of life; visit www.opswat.com.