SOLUTIONS GUIDE

UDA-REDACT

Machine Learning Powered PII Redaction for Unstructured Data — Enabling Application TDM, AI Model Training, and Agentic AI Training

Enterprise teams building AI Models, Agentic AI systems, and modern applications need access to real, representative unstructured data — PDFs, scanned forms, images — for training, testing, and validation. The challenge: this unstructured data is dense with Personally Identifiable Information — Social Security Numbers, tax identifiers, patient records, financial account data — that cannot legally or safely be used outside a controlled environment.

UDA-Redact eliminates this bottleneck. It uses a deep learning model to permanently remove PII from unstructured data — giving your application testing (TDM), AI Model training, and Agentic AI teams safe, compliant document data to work with. Once redacted, the output becomes a template to generate further unstructured data variants covering edge-cases and negative cases. Full operator control, tamper-evident audit trail, zero data egress.

UDA-Redact Capabilities Overview

UDA-Redact is part of GenRocket's Unstructured Data Offering — a purpose-built platform for detecting, reviewing, and permanently redacting PII from unstructured data at enterprise scale. It operates on PDFs and images today, with support for additional unstructured formats coming. Unlike conventional approaches that rely on rules, regular expressions, or manual review alone,

UDA-Redact uses a deep learning model trained specifically on the semantic structure of enterprise document formats — understanding not just what a field looks like, but what it means in context.

1. Detect

The machine learning model scans the document and identifies PII regions with confidence scores. No rules to write. No patterns to maintain.

2. Review & Redact

The operator reviews each detected region, confirms, rejects, or adds manual redactions. Every decision is logged. The model learns from every correction.

3. Export & Train

A pixel-level redacted PDF is saved — clean, structurally intact, and safe for AI Model and Agentic AI training pipelines. An immutable audit log records every action with SHA-256 proof. A governance CSV is available for compliance export.

For organisations in regulated industries — financial services, healthcare, insurance, and government — UDA-Redact makes it possible to safely prepare unstructured data for three critical purposes:

- Test Data Management (TDM) for application testing,
- AI Model training, and
- Agentic AI development.

The deep learning model handles detection and redaction; the clean output fuels whatever downstream pipeline needs it — and can serve as a template to generate further unstructured data variants without touching real PII.

Key Capabilities

Machine Learning Auto-Detection

A deep learning model trained on structured document formats identifies PII fields automatically — SSNs, EINs, names, addresses, account numbers, dates, and more — with per-field confidence scores. No rules engine. No regex patterns to maintain.

Human-in-the-Loop Review

Every detection is presented to the operator for confirmation or rejection before any permanent change is made. Operators can add manual redactions for fields the model did not detect. No redaction is applied without explicit approval.

Pixel-Level Permanent Redaction

Redacted content is permanently removed at the pixel level — not overlaid with a text box that can be stripped. The underlying data is irreversibly eliminated from the output file. What is removed stays removed.

Continuous Model Learning

Every manual redaction made by an operator is fed back into the machine learning model as a training signal. The model continuously improves its detection accuracy for your specific document types without requiring manual retraining.

Immutable Audit Log

Every redaction, every confirmation, every operator action is recorded in an append-only audit log with SHA-256 file hashes providing cryptographic chain-of-custody proof. The log cannot be modified or deleted.

Governance CSV Export

Export a structured governance log with one row per redacted region: entity type, source (auto or manual), detected text, confidence score, bounding box coordinates, and input/output file hashes. Built for your CISO, legal team, and auditors.

Batch Processing

Drop multiple PDFs or image files at once for concurrent processing. Download all redacted outputs as a single ZIP archive. Designed for high-volume enterprise document workflows.

100% Offline — Zero Data Egress

UDA-Redact runs entirely inside a single Docker container. No external API calls. No large language model. No network egress of any kind. Deploy on-premise, in a private cloud, or in a fully air-gapped environment.

Typical Use Cases

TDM, AI Model Training & Agentic AI Training

UDA-Redact serves three downstream use cases. First and foremost: Test Data Management (TDM) for application testing — QA and development teams need realistic unstructured data (PDFs, images) to test document-processing applications without compliance exposure; UDA-Redact makes that possible. Second: AI Model training — production unstructured data cannot be fed into AI training pipelines without removing PII; UDA-Redact's deep learning model makes them safe to use at scale. Third: Agentic AI training — Agentic systems that autonomously process documents need realistic, diverse unstructured training data to perform accurately; redacted documents provide exactly that. Once redacted, each document becomes a template: feed it into GenRocket's Unstructured Data Accelerator to generate unlimited synthetic unstructured data variants (PDFs and images today, more formats coming).

Compliance & Regulatory Review

Regulated organisations must demonstrate that PII is handled, redacted, and disposed of correctly. UDA-Redact's immutable audit log and governance CSV export provide the tamper-evident chain-of-custody records that compliance officers, legal teams, and external auditors require under GDPR, HIPAA, CCPA, and SOX.

Secure Document Sharing

Legal, HR, and finance teams frequently need to share document copies across departments or with external parties — regulators, auditors, vendors — with sensitive fields removed. UDA-Redact automates this process, replacing manual redaction workflows that are slow, error-prone, and not auditable at scale.

Application TDM & Developer Environment Provisioning

QA teams, developers, and application testers need realistic unstructured data — PDFs and images — to build, test, and validate document processing systems. UDA-Redact enables safe provisioning of redacted real documents into test and development environments, eliminating compliance risk while preserving the structural realism that testing requires. Redacted documents can also serve as templates in GenRocket's Unstructured Data Accelerator to generate high-volume synthetic unstructured TDM datasets on demand.

Industry Applications

Financial Services

Financial institutions process enormous volumes of PII-dense documents — W-2s, pay stubs, loan applications, mortgage packets, wire transfer forms, and trade confirmations. Regulations such as GDPR, CCPA, GLBA, and SOX restrict how these documents can be stored, shared, and used for testing. Manual redaction is expensive and creates inconsistent audit trails.

UDA-Redact automates PII removal across these document types, enabling financial services teams to safely prepare documents for AI Model and Agentic AI training, IDP system validation, and regulatory audit submission — without exposing customer financial data.

Example Use Case: Mortgage Processing Automation

A regional bank processes tens of thousands of mortgage applications each month. Each packet includes W-2s, pay stubs, identification cards, and bank statements containing dense PII. The compliance team needed to prepare a large document corpus for training a new document classification model — but could not use live customer files. Using UDA-Redact, they automated PII detection and redaction across the entire document corpus in hours rather than weeks, producing a clean, auditable dataset safe for model training and compliant with GLBA and SOX requirements.

Healthcare

Healthcare providers and payers handle some of the most sensitive PII under the tightest regulatory constraints — patient intake forms, insurance claims, explanation of benefits, EHR records, and laboratory reports, all governed by HIPAA and state-level privacy laws. Even anonymised datasets carry residual re-identification risk.

UDA-Redact removes PHI and PII from healthcare documents with pixel-level permanence, enabling clinical informatics teams, EHR vendors, and healthcare AI developers to work with realistic document samples without regulatory exposure.

Example Use Case: AI Model Training for Claims Adjudication

A large healthcare payer needed to train an AI Model to automate claims adjudication and build an Agentic AI system for end-to-end claims routing. Privacy regulations made it impossible to use real patient claims — developers were limited to a handful of manually sanitised samples, leading to poor coverage and inconsistent AI performance. Using UDA-Redact, the organisation processed thousands of historical claims, automatically detecting and permanently removing all PHI. The resulting corpus enabled full AI Model training coverage while maintaining complete HIPAA compliance — and the clean redacted documents became the foundation for their Agentic AI training pipeline.

Banking & Insurance

Banks and insurance carriers process complex, multi-page document packets — policy applications, claims packets, underwriting files, and customer identity documents — that combine structured tabular data with unstructured free-form content. Testing document processing pipelines with these files without exposing customer data is a persistent challenge.

UDA-Redact handles mixed-format documents, identifying PII across both structured fields and free-form text regions with high accuracy. The human-in-the-loop review step ensures that complex documents with non-standard layouts are handled correctly before any permanent redaction is applied.

Example Use Case: Insurance Claims Document Pipeline

A national insurance carrier was building an automated claims intake pipeline powered by OCR and ML-based classification. The QA team needed a large, diverse corpus of realistic claims documents to test the pipeline across edge cases — partial scans, handwritten fields, multi-page packets. Using UDA-Redact, they automated PII removal from thousands of historical claims, producing a clean test corpus. The governance log provided the compliance team with a complete audit record of every field redacted, satisfying internal data governance requirements.

Customer Challenges Solved with UDA-Redact

CHALLENGE	HOW UDA-REDACT SOLVES IT
Production unstructured data contains PII that blocks application TDM, AI Model training, Agentic AI development, and cross-team sharing.	Permanently removes PII at the pixel level, producing clean, structurally intact unstructured data safe for TDM, AI Model training, Agentic AI pipelines, and cross-team sharing.
Manual redaction workflows are slow, inconsistent, and produce no auditable record.	Automates detection and provides a structured, immutable audit log with SHA-256 chain-of-custody proof for every redaction performed.
Rules-based redaction tools miss PII in non-standard fields or document layouts.	Deep learning model trained on document structure detects PII semantically — not just by pattern — and learns from every manual correction.
Compliance teams cannot prove what was redacted, when, and by whom.	Governance CSV export provides a row-per-region record of every redaction: entity type, source, confidence score, operator, timestamp, and file hashes.
Sending documents to cloud-based redaction services creates unacceptable data sovereignty risk.	100% offline, Docker-native deployment. No external API calls, no network egress, no cloud upload. Runs entirely within your own infrastructure.
Redacted documents need to be structurally intact and usable for AI Model and Agentic AI training.	Non-PII content is fully preserved — structure, layout, and context remain intact. Redacted documents feed directly into AI training pipelines or into GenRocket's UDA Accelerator to generate synthetic training variants at scale.

Benefits Realized

- **Compliance & Privacy**: Eliminates compliance risk by replacing PII-dense production documents with permanently redacted, auditable outputs.
- **Automation at Scale**: Machine learning detection operates at speed — processing documents in seconds and scaling to thousands of files via batch mode.
- **Audit-Ready**: Every redaction is logged with cryptographic proof, satisfying GDPR, HIPAA, CCPA, SOX, and internal data governance requirements.
- **Continuous Improvement**: The model improves automatically with every operator correction — no manual retraining required.
- **AI, Agentic & TDM Ready**: Redacted unstructured data (PDFs and images today, more formats coming) is ready for TDM application testing, AI Model training, and Agentic AI training. Feed redacted documents into GenRocket's UDA Accelerator to generate unlimited synthetic unstructured data variants — no PII, full structural realism.
- **Cost Reduction**: Eliminates costly, error-prone manual redaction workflows, reducing document preparation time from weeks to hours.

How UDA-Redact Is Delivered

UDA-Redact is part of GenRocket's Unstructured Data Offering. It can be deployed as a standalone redaction platform or as the first stage of a complete document pipeline in combination with GenRocket's Unstructured Data Accelerator (UDA).

The complete pipeline — Redact then Generate — permanently removes PII from unstructured data and uses those clean documents as templates to generate hundreds of synthetic variants for TDM application testing, AI Model training, and Agentic AI training. PDFs and images are supported today, with more unstructured formats on the roadmap. Every synthetic variant is structurally identical to the real document — giving AI systems and testing pipelines the realistic, diverse unstructured data they need.

UDA-Redact is delivered as a Docker container with all machine learning models bundled. A single docker run command deploys the full platform — no GPU required, no external dependencies, no internet access needed.

If your organisation is building AI Models or Agentic AI systems that require enterprise document training data — or facing PII compliance challenges in your document workflows — contact your GenRocket account director to schedule a discovery session with our UDA-Redact specialists.