International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 5, September – October 2025 ||

DOI: 10.15680/IJCTECE.2025.0805017

"Zero Trust, Full Intelligence: PI/SPI/PHI/NPI/PCI Redaction Strategies for Agentic and Next-Gen AI Ecosystems"

Dr. Rashmiranjan Pradhan

AI, Gen AI, Agentic AI Innovation leader at IBM, Bangalore, Karnataka, India.

rashmiranjan.pradhan@gmail.com

ABSTRACT: The proliferation of Artificial Intelligence (AI), Generative AI (GenAI), and autonomous Agentic AI systems promises unprecedented analytical capabilities but simultaneously exacerbates data privacy and security risks. These systems often require access to vast datasets containing highly sensitive information, including Personal Information (PI), Sensitive Personal Information (SPI), Protected Health Information (PHI), Nonpublic Personal Information (NPI), and Payment Card Information (PCI). Traditional perimeter-based security models are inadequate for the dynamic, distributed nature of modern AI ecosystems. This paper proposes integrating Zero Trust principles with advanced data redaction strategies to create a robust framework for safeguarding sensitive data while maximizing its utility for AI. We analyze technical approaches for identifying and redacting PI, SPI, PHI, NPI, and PCI across diverse industry contexts (healthcare, finance, telecommunications). We explore the application of techniques ranging from rule-based pattern matching and Named Entity Recognition (NER) to context-aware language models and tokenization. The paper presents a conceptual Zero Trust Data Redaction Lifecycle and discusses its implementation within complex AI pipelines, including challenges specific to GenAI and Agentic AI, such as prompt injection risks and ensuring agent privacy compliance. We provide a comparative analysis of redaction methods and highlight opportunities for developing privacy-aware AI systems. The proposed framework offers a pathway to achieving "Full Intelligence" from data assets without compromising the imperative of Zero Trust data protection.

KEYWORDS: “Zero Trust,” "Data Redaction," "Artificial Intelligence," "Generative AI," "Agentic AI," "Data Privacy," "Data Security," "PI," "SPI," "PHI," "NPI," "PCI," "Privacy-Preserving AI," "GDPR," "HIPAA," "PCI DSS," "Data Protection," " IEEE Standards."

I. INTRODUCTION

The transformative potential of artificial intelligence across industries is undeniable. From diagnosing diseases and detecting financial fraud to personalizing customer experiences and automating complex tasks, AI systems are becoming integral to modern operations. However, the effectiveness of AI models, particularly those employing deep learning and large language models (LLMs), is heavily reliant on access to large, diverse, and often highly sensitive datasets. These datasets frequently contain categories of information subject to stringent regulatory requirements and ethical considerations, including Personal Information (PI), Sensitive Personal Information (SPI), Protected Health Information (PHI) governed by regulations like HIPAA, Nonpublic Personal Information (NPI) under acts such as the Gramm-Leach-Bliley Act (GLBA), and Payment Card Information (PCI) subject to standards like PCI DSS.

Simultaneously, the security landscape has evolved. The traditional security perimeter has dissolved with the rise of cloud computing, mobile workforces, and complex interconnected systems. This shift necessitates a move towards Zero Trust security models, where no entity, inside or outside the network, is implicitly trusted. Every access request is explicitly verified.

The confluence of these trends presents a critical challenge: How can organizations leverage sensitive data for AI development and deployment in a manner that is both compliant with privacy regulations (e.g., GDPR, CCPA) and aligned with Zero Trust principles, especially as AI ecosystems become more distributed and autonomous with GenAI and Agentic AI? Simply locking down data inhibits innovation, while unrestricted access is a compliance and security catastrophe waiting to happen.

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 5, September – October 2025 ||

DOI: 10.15680/IJCTECE.2025.0805017

Data redaction – the irreversible or pseudonymizing removal or masking of sensitive information – emerges as a crucial technique within a Zero Trust framework. By redacting data before it is used for AI training, inference, or agent interaction, the potential blast radius of a data breach or misuse incident is significantly reduced. However, effective redaction is challenging. It requires high accuracy to avoid leakage, context preservation where necessary for AI tasks, and scalability across diverse data formats and types of sensitive information.

This paper addresses this challenge by proposing and analyzing redaction strategies integrated within a Zero Trust architecture specifically tailored for next-generation AI ecosystems. We examine the technical nuances of redacting PI, SPI, PHI, NPI, and PCI, considering the unique characteristics and regulatory requirements of each. We analyze how Zero Trust principles can be applied to the data lifecycle within AI systems and investigate the specific opportunities and complexities introduced by GenAI and Agentic AI.

The key contributions of this paper are:

1. A comprehensive analysis of technical redaction strategies tailored for diverse sensitive data types (PI, SPI, PHI, NPI, PCI) across key industries.
2. A framework for integrating Zero Trust principles into the data lifecycle management for AI, emphasizing redaction as a core control.
3. An exploration of the specific challenges and considerations for implementing redaction within GenAI and Agentic AI systems.
4. A comparative overview of redaction techniques based on their applicability, effectiveness, and trade-offs in the context of AI utility.

II. RELATED WORK

Traditional approaches to data privacy often focus on anonymization techniques like k-anonymity, l-diversity, and t-closeness. These methods aim to modify datasets so that individual records cannot be uniquely identified or linked back to individuals. While valuable, these techniques can sometimes overly generalize or distort data, reducing its utility for complex AI tasks, especially those requiring fine-grained detail or natural language understanding. Furthermore, re-identification risks persist, particularly with external data sources or side-channel attacks.

Redaction, in the context of textual or structured data, is a more direct method of removing or masking specific sensitive identifiers. Early methods relied heavily on rule-based pattern matching (e.g., regular expressions for social security numbers or email addresses). The advent of Natural Language Processing (NLP) techniques, particularly Named Entity Recognition (NER), significantly improved the ability to identify and redact entities like names, locations, dates, and organizations in unstructured text. More recent advancements leveraging transformer models have further enhanced context-aware PII/PHI detection, crucial for disambiguating sensitive information in complex narratives like clinical notes.

Privacy-preserving machine learning (PPML) explores techniques like Federated Learning (training models on decentralized data without sharing raw data) and Secure Multi-Party Computation (SMPC) or Homomorphic Encryption (performing computations on encrypted data). While powerful, these methods can be computationally intensive or require significant architectural changes, and they don't always eliminate the need to handle sensitive data *at rest* or *at the ingestion point*, where redaction is still applicable. Differential Privacy offers strong, quantifiable privacy guarantees by injecting noise into data or query results, making it difficult to infer information about any single individual. It is highly relevant for training data aggregation or releasing aggregate statistics, but direct redaction is often necessary for compliance mandates regarding specific data element removal.

Zero Trust security architectures emphasize "never trust, always verify." This model shifts focus from network perimeter defense to protecting resources (including data) by strictly controlling access based on identity, context, and policy. While Zero Trust frameworks are well-established for network and application access, their specific application to *data processing pipelines* within complex AI ecosystems, particularly integrating data transformation techniques like redaction, requires further exploration.

Regulatory frameworks like GDPR, CCPA, HIPAA, and PCI DSS mandate specific requirements for handling sensitive data, often including the right to erasure or the necessity of masking certain data elements for compliance and risk reduction. Redaction is a direct technical control to help meet these requirements.

International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 5, September – October 2025 ||

DOI: 10.15680/IJCTECE.2025.0805017

Existing work has explored PII detection and anonymization for specific domains like healthcare or legal text. However, a unified approach that integrates Zero Trust principles with sophisticated redaction strategies across diverse sensitive data types (PI, SPI, PHI, NPI, PCI) and specifically addresses the unique demands and risks of next-generation AI, GenAI, and Agentic AI ecosystems is an area requiring deeper analysis. This paper aims to bridge this gap by proposing such a framework and analyzing its practical implementation.

III. METHODOLOGY: ZERO TRUST DATA REDACTION LIFECYCLE FOR AI

Integrating Zero Trust with data redaction for AI requires a lifecycle approach that enforces verification and least privilege at every stage where sensitive data is handled or processed by AI components. We propose the Zero Trust Data Redaction Lifecycle (ZTDRL) framework, illustrated conceptually in Figure 1 (Note: Figure 1 is a conceptual placeholder).

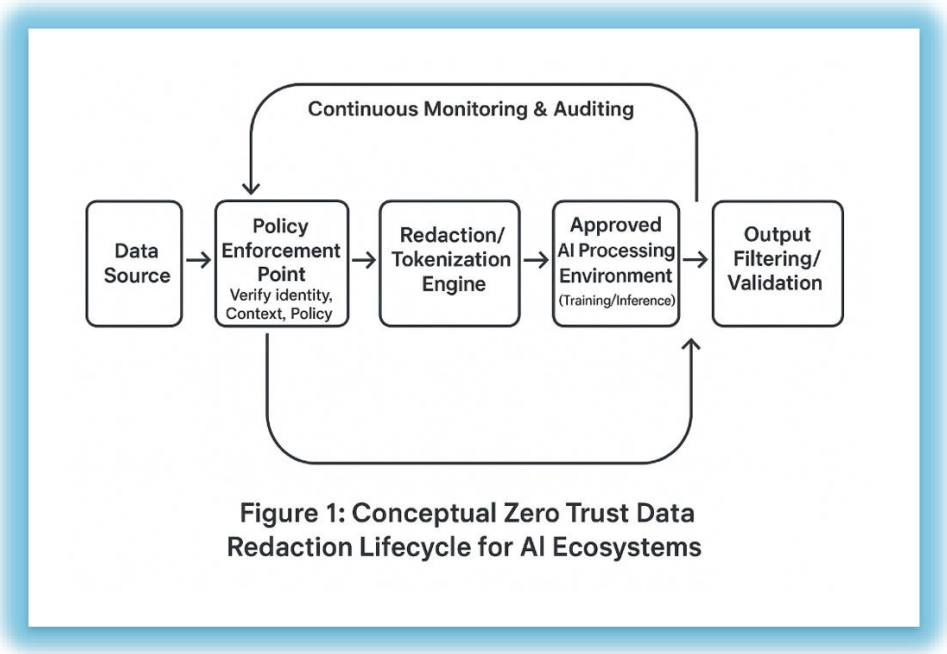

Figure 1: Conceptual Zero Trust Data Redaction Lifecycle for AI Ecosystems

The ZTDRL comprises the following stages:

1. Secure Data Ingestion and Classification: Data from various sources (databases, data lakes, APIs, documents) is ingested into a secure landing zone. Automated tools are used to identify and classify sensitive data types (PI, SPI, PHI, NPI, PCI) based on patterns, dictionaries, and machine learning models. This initial classification informs access policies.
2. Zero Trust Policy Enforcement Point (PEP): Any request to access or process the ingested data, whether by an AI training pipeline, an inference service, a GenAI model, or an Agentic AI component, must first pass through a PEP. The PEP verifies:
    - Requester Identity: Strong authentication of the AI service, user, or agent.
    - Context: The purpose of access (training, inference, analytics), the source/destination system, time of access, and perceived risk.
    - Policy: Evaluate access rules based on verified identity, context, and the sensitivity classification of the data. Policies define what data can be accessed and in what state (e.g., raw, redacted, tokenized) for what purpose. Least privilege is strictly enforced – components only get the minimum data necessary.
3. Dynamic Redaction/Tokenization Engine: If the policy allows access only to a redacted or tokenized version of the data, the request is routed to a dedicated Redaction/Tokenization Engine. This engine applies the appropriate technique (discussed in Section IV) based on the data type and the policy mandate. This process is dynamic, meaning redaction is applied on demand based on the authorized request, rather than necessarily storing only redacted copies (though static redaction for training sets is common).

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 5, September – October 2025 ||

DOI: 10.15680/IJCTECE.2025.0805017

4. Approved AI Processing Environment: The AI system (training cluster, inference endpoint, GenAI service, agent runtime) receives the redacted or tokenized data. It operates within a secure, isolated environment where access to external resources is strictly controlled. The AI model processes the data to perform its intended task.
5. Output Filtering and Validation: AI model outputs, especially from GenAI or Agentic systems, must be filtered and validated to prevent the unintentional leakage or regeneration of sensitive information. This involves applying similar redaction or pattern-matching techniques to the output before it is delivered to the end-user or downstream system.
6. Continuous Monitoring and Auditing: All access requests, policy decisions, redaction events, and data flows are logged and monitored. Anomalies or policy violations trigger alerts and automated responses. This provides traceability and accountability, essential components of a Zero Trust architecture and compliance requirements.

This ZTDRL ensures that sensitive data, even when residing within the environment, is not inherently trusted and is only accessed in a de-sensitized form for AI processing based on explicit verification and authorization.

IV. REDACTION STRATEGIES AND TECHNOLOGIES FOR PI/SPI/PHI/NPI/PCI

Effective redaction requires a nuanced approach, considering the specific characteristics of each data category and the context in which it appears.

A. Personal Information (PI) and Sensitive Personal Information (SPI)

PI (e.g., name, address, email, phone number) and SPI (e.g., SSN, driver's license, race, religion, biometric data) are pervasive across all industries.

- **Techniques:**
  - *Rule-Based:* Regular expressions are effective for structured formats like SSNs (e.g., ^\d{3}-\d{2}-\d{4}$), email addresses, and phone numbers. Dictionaries can be used for common names or locations, though prone to false positives/negatives.
  - *Machine Learning (NER):* Deep learning models, often based on transformers (e.g., BERT, RoBERTa), are highly effective for identifying names, organizations, locations, dates, and other entities in unstructured text, providing better context awareness than rules alone.
  - *Substitution/Masking:* Replacing identified PI/SPI with a placeholder (e.g., "[PERSON]", "[EMAIL]") or masking partial data (e.g., XXX-XX-XXXX for SSN).
- **Technologies:** NLP libraries (spaCy, NLTK, Hugging Face Transformers), cloud provider PII detection APIs (AWS Comprehend, Google Cloud Natural Language, Azure Text Analytics), commercial data loss prevention (DLP) tools.

B. Protected Health Information (PHI)

PHI, as defined by HIPAA, includes demographic information, medical history, test results, insurance information, and other data linked to an individual's health status. PHI is particularly challenging due to its often unstructured nature (clinical notes, doctor's reports) and the need to identify relationships and context.

- **Techniques:**
  - *Specialized NER:* Requires models trained specifically on clinical text, as medical terminology and abbreviations are unique. Identifying PHI like patient names, doctor names, hospital names, dates of service, medical record numbers, and body parts requires domain expertise.
  - *Contextual Analysis:* Distinguishing between a patient's name and a doctor's name appearing in the same note requires understanding sentence structure and context. Transformer models excel here.
  - *De-identification Safe Harbor/Expert Determination:* While redaction is a technical control, achieving HIPAA de-identification status may require removing specific identifiers listed in the Safe Harbor (e.g., all dates except year, geographic subdivisions smaller than a state) or obtaining expert certification. Redaction tools aim to facilitate these requirements.
- **Technologies:** Specialized healthcare NLP APIs (AWS Comprehend Medical, Google Healthcare API), academic tools (e.g., scispaCy), commercial PHI de-identification software.

C. Nonpublic Personal Information (NPI)

NPI in finance includes data provided by a consumer to a financial institution (e.g., account balances, transaction history, creditworthiness, loan information).

- **Techniques:**
  - *Rule-Based/Pattern Matching:* Effective for account numbers (following specific bank formats), policy numbers, and potentially transaction IDs.

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 5, September – October 2025 ||

DOI: 10.15680/IJCTECE.2025.0805017

- *Entity Matching:* Identifying customer names, addresses, or related parties involved in financial transactions, often linked to internal customer databases.
- *Masking/Substitution:* Masking parts of account numbers or replacing customer names with internal IDs.
- *Tokenization:* Replacing sensitive NPI values (like account numbers) with unique, non-sensitive tokens. This is often reversible via a secure token vault for specific authorized processes, but for AI analysis, non-reversible or utility-preserving tokenization is preferred.

- **Technologies:** DLP systems, data masking tools, specialized financial data tokenization platforms, database security features.

D. Payment Card Information (PCI)

PCI includes the full Primary Account Number (PAN), cardholder name, expiration date, and Service Code. The Card Security Code (CSC) or CVV is also highly sensitive but *must not* be stored after authorization, so redaction focuses on PAN and associated data.

- **Techniques:**
  - *Pattern Matching:* Luhn algorithm and specific issuer ranges for PAN detection.
  - *Masking:* Displaying only the first few and last four digits of the PAN (e.g., 4111XXXXXXXX1111). This is a common requirement for display but often insufficient for AI processing if the AI needs to analyze patterns across full PANs (in which case tokenization is better).
  - *Tokenization:* Replacing the PAN with a unique token that can be used for downstream processing without exposing the actual card number. This is the *de facto* standard for handling PCI in payment systems and increasingly used for analytics.
- **Technologies:** PCI-certified tokenization platforms, hardware security modules (HSMs) for key management, secure payment gateways that handle tokenization. Compliance with PCI DSS is paramount.

V. INTEGRATING REDACTION WITH AI, GENAI, AND AGENTIC AI ECOSYSTEMS

The application of redaction within AI ecosystems necessitates integration into data pipelines and AI model workflows.

A. Traditional AI (Supervised/Unsupervised Learning)

For training traditional models, sensitive data should be redacted *before* being included in the training dataset. This minimizes the risk of the model memorizing or leaking sensitive training examples. During inference, input data containing sensitive information must be redacted *before* being fed into the model, and the model's output should be scanned for potential leakage. This is typically implemented via pre-processing and post-processing steps in the data pipeline or inference service.

B. Generative AI (GenAI)

GenAI models, particularly LLMs, pose unique challenges due to their ability to generate novel text based on learned patterns.

- **Prompt Sanitization:** User prompts fed to a GenAI model must be scanned and redacted if they contain sensitive information. This prevents the model from being exposed to PI/SPI/PHI/NPI/PCI directly via the input.
- **Output Filtering:** The output generated by the GenAI model must be scanned and redacted *before* being presented to the user or used downstream. LLMs can sometimes regurgitate training data or infer sensitive details from context, even if the input prompt was clean.
- **Fine-tuning Data Redaction:** If fine-tuning GenAI models on domain-specific data (e.g., customer support logs, internal documents), the fine-tuning dataset must be rigorously redacted.
- **Zero Trust Principle Application:** Applying "Verify Explicitly" means the GenAI service identity is verified. "Least Privilege" dictates that the service only interacts with the redaction engine and allowed data sources. "Assume Breach" means both input prompts and model outputs are treated as potentially containing sensitive data and must be filtered.

C. Agentic AI

Agentic AI systems involve autonomous or semi-autonomous agents that can interact with their environment, call tools, and make decisions based on goals. This introduces further complexity for redaction and Zero Trust.

- **Agent Identity and Authorization:** Each agent needs a strong identity, and its actions, including accessing data sources or calling APIs (like a redaction service or a sensitive database), must be explicitly authorized based on Zero Trust policies.

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 5, September – October 2025 ||

DOI: 10.15680/IJCTECE.2025.0805017

- **Tool Interaction:** Agents might use tools that handle sensitive data (e.g., a database query tool, an email sending tool). All data passed to and from these tools via the agent must pass through the ZTDRL PEP and Redaction Engine if required by policy.
- **Context Management:** Agents maintain context across interactions. Ensuring sensitive information is not inadvertently carried over in the agent's memory or state, even if originating from a redacted source or previously filtered output, is critical.
- **Redaction-Aware Agents:** Future Agentic AI design could incorporate privacy awareness directly into the agent's reasoning process, prompting it to request redacted data versions or utilize redaction tools proactively.
- **Challenges:** Agents' autonomous nature makes predicting their exact data access patterns difficult. Prompt injection attacks could target agents to coerce them into bypassing redaction or leaking information through tool use. Designing policies that are flexible enough for agents' dynamic behavior yet strict enough for Zero Trust is complex.

VI. INDUSTRY CASE STUDIES (CONCEPTUAL)

A. Healthcare: Training a Clinical Note Analysis Model

- **Scenario:** A hospital wants to train an NLP model to identify patient symptoms, diagnoses, and treatments from doctor's clinical notes to improve coding and research. Notes contain extensive PHI.
- **ZTDRL Application:**
  1. Notes ingested into secure data lake. Automated PHI detection classifies entities (patient name, doctor name, dates, hospital, medical record number).
  2. AI training service requests access to clinical notes for training. PEP verifies the service identity, purpose (training), and policy (access allowed only to de-identified data).
  3. Redaction Engine applies PHI-specific rules and specialized NER models to redact all 18 HIPAA identifiers, replacing them with placeholders ([PATIENT], [DATE], [HOSPITAL], etc.). Dates might be shifted or generalized to the year per Safe Harbor.
  4. Redacted notes are used for model training in an isolated environment.
  5. During inference (e.g., processing new notes for coding), input notes are redacted, and the model output (identified entities) is validated to ensure no unredacted PHi is present.
  6. All access and redaction actions are logged for HIPAA compliance audits.
- **Challenges:** Maintaining clinical context after redaction, accurately identifying all PHI in diverse note formats and language styles.

B. Finance: Transaction Monitoring for Fraud Detection

- **Scenario:** A bank uses AI to analyze customer transaction logs to detect fraudulent activity. Logs contain NPI (account numbers, transaction details) and PCI (PANs).
- **ZTDRL Application:**
  1. Transaction logs ingested into a secure platform. Automated processes identify NPI and PCI.
  2. Fraud detection AI service requests access to transaction data. PEP verifies service identity and policy (access allowed to tokenized PCI and masked NPI).
  3. Tokenization service replaces PANs with non-reversible tokens. Redaction engine masks account numbers (e.g., showing last 4 digits) and potentially substitutes sensitive merchant names with categories.
  4. AI analyzes tokenized/masked data for patterns indicative of fraud. The model operates on tokens, which are meaningless outside the tokenization system.
  5. AI alerts generated based on patterns in tokenized data. If an alert requires investigating the original transaction, a separate, highly restricted process with explicit multi-factor authentication retrieves the original data, bypassing standard AI access.
  6. Auditing tracks all access to both tokenized and original data.
- **Challenges:** Balancing data utility for complex fraud patterns with strict PCI and NPI requirements, managing the tokenization service securely.

C. Telecommunications: Customer Service Chatbot Improvement

- **Scenario:** A telecom company uses customer interaction logs from chatbots and support calls (transcribed) to train and improve its customer service AI, including a GenAI chatbot. Logs contain PI (name, address, phone), SPI (account PIN, partial SSN), and sometimes NPI (billing details).
- **ZTDRL Application:**

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 5, September – October 2025 ||

DOI: 10.15680/IJCTECE.2025.0805017

1. Interaction logs are ingested. Automated tools identify PI/SPI/NPI.
2. AI training pipeline for language models requests access. PEP verifies identity and policy (access only to redacted logs).
3. Redaction engine applies rules and NER to remove names, addresses, phone numbers, mask PINs/SSNs, and generalize billing specifics.
4. Redacted logs train the AI model.
5. For the live GenAI chatbot:
    - Input prompts are scanned and redacted in real-time before reaching the LLM.
    - LLM output is scanned and redacted before being sent to the user.
6. An Agentic AI component interacting with customer accounts would require explicit authorization via the PEP for every attempt to retrieve customer details, and the retrieved data would be redacted before the agent processes it. The agent's goal-driven actions involving sensitive data are strictly policed.

- **Challenges:** Real-time redaction latency for live interactions, preventing LLM hallucination of sensitive data, managing context for agents while redacting identifiers.

VII. OPPORTUNITIES AND CHALLENGES IN AGENTIC AI

Agentic AI systems represent a significant step towards autonomous action, bringing both opportunities and magnified challenges for sensitive data handling.

Opportunities:

- **Privacy Enforcement Agents:** Agents could be designed specifically to act as guardians, enforcing redaction policies and monitoring data flows between other agents or systems.
- **Context-Aware Redaction Requests:** Agents could potentially understand the *purpose* of their data need and request the *minimally required redacted form* of the data, moving beyond simple static redaction.
- **Automated Remediation:** Agents could potentially be tasked with identifying and automatically redacting sensitive data in specified data stores based on defined policies.

Challenges:

- **Policy Complexity:** Defining Zero Trust policies for dynamic, interacting agents is far more complex than for static services or user access. Policies must account for agent goals, tools, and potential execution paths.
- **Trusting the Agent:** While the system verifies the *agent's identity*, can we fully trust the agent's *intent* or *reasoning process* regarding sensitive data, especially with opaque large models driving agent behavior?
- **Prompt Injection and Exfiltration:** Malicious actors could craft prompts or manipulate tools to trick an agent into bypassing redaction, revealing sensitive information through its actions or generated output.
- **State Management:** Agents maintain internal state and memory. Ensuring sensitive data, even if initially redacted, doesn't persist in a way that could be later reconstructed or inferred by the agent is difficult.
- **Auditing and Explainability:** Tracing an agent's decision-making process that led to accessing or handling sensitive data, and verifying that redaction was correctly applied, adds complexity to auditing and requires greater agent explainability.

Addressing these challenges requires advancements in agent design, robust verification mechanisms for agent actions, and sophisticated policy engines capable of managing dynamic access requests in real-time based on rich context.

Comparative Analysis of Redaction Approaches

Redaction techniques vary in their effectiveness, performance, and suitability depending on the data type and AI task requirements.

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 5, September – October 2025 ||

DOI: 10.15680/IJCTECE.2025.0805017

Feature	Rule-Based Pattern Matching	Machine Learning (NER)	Tokenization	Differential Privacy (as related)
Data Types	Structured, specific formats (SSN, CC#)	Unstructured text, named entities	Structured identifiers (CC#, Acct#)	Aggregate data, statistical output
Accuracy	High for strict patterns; Low for flexible/contextual	High for trained entities; Requires domain data	Very high for specific values	Guarantees aggregate privacy; Trade-off with accuracy
Context Aware	Low	High (with advanced models)	Low	N/A
Reversibility	Generally irreversible	Irreversible (typically substitution)	Can be reversible (vaulted) or irreversible	Irreversible
Performance	Fast	Moderate to Slow (model dependent)	Fast (lookup based)	Can add significant overhead
Implementation	Simple rules, regex libraries	ML models, training data, compute	Secure vault system, API calls	Complex algorithms, careful tuning
Compliance Fit	Good for specific identifiers	Good for textual PII/PHI/NPI	Essential for PCI, NPI	Good for aggregate data analysis, not specific element removal
AI Utility	Preserves structure; loses specific values	Preserves most text structure; loses named entities	Preserves structure; loses specific values (unless utility-preserving)	Perturbs data; affects model accuracy

No single method is universally superior. A layered approach combining these techniques within the ZTDRL framework is often necessary. For instance, using pattern matching for obvious identifiers, NER for contextual entities in text, and tokenization for financial data, all governed by Zero Trust policies, provides the most robust protection. The choice depends on the data type, regulatory requirements, and the specific needs of the AI application (e.g., an AI analyzing financial trends might need tokenization for PCI but masked account numbers, while a clinical NLP model needs detailed PHI redaction).

VIII. DISCUSSION

Integrating Zero Trust principles with advanced data redaction strategies offers a robust framework for secure and compliant AI ecosystems handling sensitive information, positioning redaction as a crucial enforcement point within a "never trust, always verify" data access model. While technical feasibility has improved significantly with advancements in NLP and specialized models, a persistent challenge lies in balancing the imperative for privacy with the need to preserve sufficient data utility for complex AI tasks; inaccurate redaction risks either rendering data unusable through over-removal or leaving organizations vulnerable to leaks through under-removal.

The emergence of Generative and Agentic AI introduces magnified risks due to their generative capabilities and increasing autonomy, including the potential for accidental or malicious reconstruction and leakage of sensitive data. Addressing these threats requires rigorous input sanitization, output filtering, and stringent policy enforcement at every interaction point, where the Zero Trust model proves inherently valuable by limiting AI access solely to de-sensitized information based on explicit authorization. Future research must focus on developing more context-aware and adaptable redaction methods to minimize utility loss, establishing formal verification techniques for pipeline effectiveness, proactively embedding privacy considerations into autonomous agent designs, and mitigating potential biases introduced by the redaction process.

IX. CONCLUSION

The increasing reliance on AI, GenAI, and Agentic AI necessitates a fundamental shift in how organizations protect sensitive data. Traditional perimeter defenses are obsolete in complex, distributed AI environments. This paper advocates for the adoption of a Zero Trust security model as the foundation for sensitive data protection in these next-generation ecosystems.

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 5, September – October 2025 ||

DOI: 10.15680/IJCTECE.2025.0805017

Within this framework, advanced data redaction strategies for PI, SPI, PHI, NPI, and PCI are not merely compliance checkboxes but essential technical controls. We have presented a Zero Trust Data Redaction Lifecycle methodology and analyzed specific techniques and technologies required to identify and redact diverse sensitive data types across healthcare, finance, and telecommunications. The paper highlighted the unique challenges and opportunities presented by integrating redaction within GenAI and the emerging field of Agentic AI, emphasizing the need for robust input/output filtering and stringent, context-aware access policies.

By implementing Zero Trust-integrated redaction strategies, organizations can significantly reduce the risk associated with leveraging sensitive data for AI, comply with evolving privacy regulations, and build trust with their users and customers. Achieving "Full Intelligence" from data assets is possible, but only when built upon the bedrock of "Zero Trust" security principles, with redaction serving as a vital layer of defense. This research provides a framework and analysis to guide the development of secure, compliant, and powerful AI systems in the age of pervasive data sensitivity.

REFERENCES

[1] Rawal, A., McCoy, J., Rawat, D.B., Sadler, B.M. and Amant, R.S., 2021. Recent advances in trustworthy explainable artificial intelligence: Status, challenges, and perspectives. IEEE Transactions on Artificial Intelligence, 3(6), pp.852-866.

[2] Pradhan, D. R. (2025) "Generative Agents at Scale: A Practical Guide to Migrating from Dialog Trees to LLM Frameworks," International Journal of Computer Technology and Electronics Communication (IJCTEC) . International Journal of Computer Technology and Electronics Communication (IJCTEC), 8(5), p. 11367. doi: 10.15680/IJCTECE.2025.0805010. https://ijctece.com/index.php/IJCTEC/article/view/230/192

[3] Pradhan, Dr. Rashmiranjan. "Generative Agents at Scale: A Practical Guide to Migrating from Dialog Trees to LLM Frameworks." International Journal of Computer Technology and Electronics Communication (IJCTEC) , vol. 8, no. 5, International Journal of Computer Technology and Electronics Communication (IJCTEC), 2025, p. 11367.

[4] Pradhan, D. R. (2025) "Establishing Comprehensive Guardrails for Digital Virtual Agents: A Holistic Framework for Contextual Understanding, Response Quality, Adaptability, and Secure Engagement," International Journal of Innovative Research in Computer and Communication Engineering.doi:10.15680/IJIRCCE.2025.1307013. https://ijircce.com/admin/main/storage/app/pdf/e9xlTkp5RqODN3RmJOT2uK5biLYlwDggGH9ngoi6.pdf

[5] Pradhan, D. R. (2025). Establishing Comprehensive Guardrails for Digital Virtual Agents: A Holistic Framework for Contextual Understanding, Response Quality, Adaptability, and Secure Engagement. International Journal of Innovative Research in Computer and Communication Engineering. https://doi.org/10.15680/IJIRCCE.2025.1307013

[6] Pradhan DR. Establishing Comprehensive Guardrails for Digital Virtual Agents: A Holistic Framework for Contextual Understanding, Response Quality, Adaptability, and Secure Engagement. International Journal of Innovative Research in Computer and Communication Engineering. 2025; doi:10.15680/IJIRCCE.2025.1307013

[7] Pradhan, Dr. Rashmiranjan. "Establishing Comprehensive Guardrails for Digital Virtual Agents: A Holistic Framework for Contextual Understanding, Response Quality, Adaptability, and Secure Engagement." International Journal of Innovative Research in Computer and Communication Engineering, 2025. doi:10.15680/IJIRCCE.2025.1307013.

[8] Pradhan, D. R. RAGEvalX: An Extended Framework for Measuring Core Accuracy, Context Integrity, Robustness, and Practical Statistics in RAG Pipelines. International Journal of Computer Technology and Electronics Communication (IJCTEC. https://doi.org/10.15680/IJCTECE.2025.0805001

[9] Pradhan, D. R. (2025). RAG vs. Fine-Tuning vs. Prompt Engineering: A Comparative Analysis for Optimizing AI Models. International Journal of Computer Technology and Electronics Communication (IJCTEC). https://doi.org/10.15680/IJCTECE.2025.0805004 https://ijctece.com/index.php/IJCTEC/article/view/170/132

[10] Pradhan, Rashmiranjan, and Geeta Tomar. "AN ANALYSIS OF SMART HEALTHCARE MANAGEMENT USING ARTIFICIAL INTELLIGENCE AND INTERNET OF THINGS.". Volume 54, Issue 5, 2022 (ISSN: 0367-6234). Article history: Received 19 November 2022, Revised 08 December 2022, Accepted 22 December 2022. Harbin Gongye Daxue Xuebao/Journal of Harbin Institute of Technology. https://www.researchgate.net/profile/Rashmiranjan-Pradhan/publication/384145167_Published_Scopus_1st_journal_AN_ANALYSIS_OF_SMART_HEALTHCARE_MANAGEMENT_USING_ARTIFICIAL_INTELLIGENCE_AND_INTERNET_OF_THINGS_BY_RASHMIRANJAN_PRADHAN/links/66ec21c46b101f6fa4f0f183/Published-Scopus-1st-journal-AN-ANALYSIS-OF-SMART-HEALTHCARE-MANAGEMENT-USING-ARTIFICIAL-INTELLIGENCE-AND-INTERNET-OF-THINGS-BY-RASHMIRANJAN-PRADHAN.pdf

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 5, September – October 2025 ||

DOI: 10.15680/IJCTECE.2025.0805017

[11] Pradhan, Rashmiranjan. "AI Guardian- Security, Observability & Risk in Multi-Agent Systems." International Journal of Innovative Research in Computer and Communication Engineering, 2025. doi:10.15680/IJIRCCE.2025.1305043. https://ijircce.com/admin/main/storage/app/pdf/Mff2agMyMUfCqUV9pQSD0xsLF5dCRct45mHjvt2I.pdf

[12] Pradhan, D. R. (no date) "RAGEvalX: An Extended Framework for Measuring Core Accuracy, Context Integrity, Robustness, and Practical Statistics in RAG Pipelines," International Journal of Computer Technology and Electronics Communication (IJCTEC. doi: 10.15680/IJCTECE.2025.0805001. https://ijctece.com/index.php/IJCTEC/article/view/170/132

[13] Rashmiranjan, Pradhan Dr. "Empirical analysis of agentic ai design patterns in real-world applications." (2025). https://ijircce.com/admin/main/storage/app/pdf/7jX1p7s5bDCnn971YfaAVmVcZcod52Nq76QMyTSR.pdf

[14] Pradhan, Rashmiranjan, and Geeta Tomar. "IOT BASED HEALTHCARE MODEL USING ARTIFICIAL INTELLIGENT ALGORITHM FOR PATIENT CARE." NeuroQuantology 20.11 (2022): 8699-8709. https://ijircce.com/admin/main/storage/app/pdf/7jX1p7s5bDCnn971YfaAVmVcZcod52Nq76QMyTSR.pdf

[15] Rashmiranjan, Pradhan. "Contextual Transparency: A Framework for Reporting AI, Genai, and Agentic System Deployments across Industries." (2025). https://ijircce.com/admin/main/storage/app/pdf/OUmQRqDgcqyYJ9jHFHGVpo0qIvpQNBV9cNihzyjz.pdf

[16] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557-570, Oct. 2002.

[17] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, pp. 3-es, Mar. 2007.

[18] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond L-diversity," in *Proc. 23rd Int. Conf. Data Engineering (ICDE)*, Istanbul, Turkey, Apr. 2007, pp. 106-115.

[19] F. Finkel, S. Bethard, and K. Jurafsky, "Joint parsing and named entity recognition," in *Proc. Human Language Technology Conf. of the NAACL*, New York, NY, USA, Jun. 2005, pp. 188-195.

[20] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[21] B. Wu, S. Liu, S. Wei, R. Xu, and J. Xu, "A review of detecting protected health information (PHI) in clinical text," *Journal of Biomedical Informatics*, vol. 80, pp. 69-80, Apr. 2018.

[22] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. y Arcas, "Communication-efficient learning of deep networks from decentralized data," *arXiv preprint arXiv:1602.05629*, 2016.

[23] Y. Lindell, "Secure multiparty computation (MPC)," *Foundations and Trends® in Theoretical Computer Science*, vol. 6, no. 1–2, pp. 1-224, 2017.

[24] C. Gentry, "Fully homomorphic encryption using ideal lattices," in *Proc. 41st Annu. ACM Symp. Theory of Computing*, Bethesda, MD, USA, May-Jun. 2009, pp. 169-178.

[25] C. Dwork, "Differential privacy," in *Proc. 33rd Int. Colloquium on Automata, Languages, and Programming (ICALP)*, Venice, Italy, Jul. 2006, pp. 1-12.

[26] J. Kindervag, J. Hauck, and D. Walden, "No More Chewy Centers: Introducing The Zero Trust Model of Information Security," Forrester Research, Inc., Cambridge, MA, USA, 2010.

[27] European Union, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016," *Official Journal of the European Union*, 2016.

[28] California Legislative Information, "California Consumer Privacy Act (CCPA)," Assembly Bill No. 375, 2018.

[29] U.S. Department of Health & Human Services, "Summary of the HIPAA Privacy Rule," 45 CFR § 164.501 et seq., Apr. 2013. [Online]. Available: https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html

[30] Payment Card Industry Security Standards Council, *Payment Card Industry Data Security Standard (PCI DSS) v4.0*, Mar. 2022.

[31] Pradhan, D. R. (2025) "Generative Agents at Scale: A Practical Guide to Migrating from Dialog Trees to LLM Frameworks," International Journal of Computer Technology and Electronics Communication (IJCTEC) . International Journal of Computer Technology and Electronics Communication (IJCTEC), 8(5), p. 11367. doi: 10.15680/IJCTECE.2025.0805010. https://ijctece.com/index.php/IJCTEC/article/view/230/192

[32] Pradhan, Dr. Rashmiranjan. "Generative Agents at Scale: A Practical Guide to Migrating from Dialog Trees to LLM Frameworks." International Journal of Computer Technology and Electronics Communication (IJCTEC) , vol. 8, no. 5, International Journal of Computer Technology and Electronics Communication (IJCTEC), 2025, p. 11367.

[33] Pradhan, D. R. (2025) "Establishing Comprehensive Guardrails for Digital Virtual Agents: A Holistic Framework for Contextual Understanding, Response Quality, Adaptability, and Secure Engagement," International Journal of Innovative Research in Computer and Communication Engineering.doi:10.15680/IJIRCCE.2025.1307013. https://ijircce.com/admin/main/storage/app/pdf/e9xlTkp5RqODN3RmJOT2uK5biLYlwDggGH9ngoi6.pdf

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 5, September – October 2025 ||

DOI: 10.15680/IJCTECE.2025.0805017

[34] Pradhan, D. R. (2025). Establishing Comprehensive Guardrails for Digital Virtual Agents: A Holistic Framework for Contextual Understanding, Response Quality, Adaptability, and Secure Engagement. International Journal of Innovative Research in Computer and Communication Engineering. https://doi.org/10.15680/IJIRCCE.2025.1307013

[35] Pradhan DR. Establishing Comprehensive Guardrails for Digital Virtual Agents: A Holistic Framework for Contextual Understanding, Response Quality, Adaptability, and Secure Engagement. International Journal of Innovative Research in Computer and Communication Engineering. 2025; doi:10.15680/IJIRCCE.2025.1307013

[36] Pradhan, Dr. Rashmiranjan. "Establishing Comprehensive Guardrails for Digital Virtual Agents: A Holistic Framework for Contextual Understanding, Response Quality, Adaptability, and Secure Engagement." International Journal of Innovative Research in Computer and Communication Engineering, 2025. doi:10.15680/IJIRCCE.2025.1307013.

[37] Pradhan, D. R. RAGEvalX: An Extended Framework for Measuring Core Accuracy, Context Integrity, Robustness, and Practical Statistics in RAG Pipelines. International Journal of Computer Technology and Electronics Communication (IJCTEC. https://doi.org/10.15680/IJCTECE.2025.0805001

[38] Pradhan, D. R. (2025). RAG vs. Fine-Tuning vs. Prompt Engineering: A Comparative Analysis for Optimizing AI Models. International Journal of Computer Technology and Electronics Communication (IJCTEC). https://doi.org/10.15680/IJCTECE.2025.0805004 https://ijctece.com/index.php/IJCTEC/article/view/170/132

[39] Pradhan, Rashmiranjan, and Geeta Tomar. "AN ANALYSIS OF SMART HEALTHCARE MANAGEMENT USING ARTIFICIAL INTELLIGENCE AND INTERNET OF THINGS.". Volume 54, Issue 5, 2022 (ISSN: 0367-6234). Article history: Received 19 November 2022, Revised 08 December 2022, Accepted 22 December 2022. Harbin Gongye Daxue Xuebao/Journal of Harbin Institute of Technology. https://www.researchgate.net/profile/Rashmiranjan-Pradhan/publication/384145167_Published_Scopus_1st_journal_AN_ANALYSIS_OF_SMART_HEALTHCARE_MANAGEMENT_USING_ARTIFICIAL_INTELLIGENCE_AND_INTERNET_OF_THINGS_BY_RASHMIRANJAN_PRADHAN/links/66ec21c46b101f6fa4f0f183/Published-Scopus-1st-journal-AN-ANALYSIS-OF-SMART-HEALTHCARE-MANAGEMENT-USING-ARTIFICIAL-INTELLIGENCE-AND-INTERNET-OF-THINGS-BY-RASHMIRANJAN-PRADHAN.pdf

[40] Pradhan, Rashmiranjan. "AI Guardian- Security, Observability & Risk in Multi-Agent Systems." International Journal of Innovative Research in Computer and Communication Engineering, 2025. doi:10.15680/IJIRCCE.2025.1305043. https://ijircce.com/admin/main/storage/app/pdf/Mff2agMyMUfCqUV9pQSD0xsLF5dCRct45mHjvt2I.pdf

[41] Pradhan, D. R. (no date) "RAGEvalX: An Extended Framework for Measuring Core Accuracy, Context Integrity, Robustness, and Practical Statistics in RAG Pipelines," International Journal of Computer Technology and Electronics Communication (IJCTEC. doi: 10.15680/IJCTECE.2025.0805001. https://ijctece.com/index.php/IJCTEC/article/view/170/132

[42] Rashmiranjan, Pradhan Dr. "Empirical analysis of agentic ai design patterns in real-world applications." (2025). https://ijircce.com/admin/main/storage/app/pdf/7jX1p7s5bDCnn971YfaAVmVcZcod52Nq76QMyTSR.pdf

[43] Pradhan, Rashmiranjan, and Geeta Tomar. "IOT BASED HEALTHCARE MODEL USING ARTIFICIAL INTELLIGENT ALGORITHM FOR PATIENT CARE." NeuroQuantology 20.11 (2022): 8699-8709. https://ijircce.com/admin/main/storage/app/pdf/7jX1p7s5bDCnn971YfaAVmVcZcod52Nq76QMyTSR.pdf

[44] Rashmiranjan, Pradhan. "Contextual Transparency: A Framework for Reporting AI, Genai, and Agentic System Deployments across Industries." (2025). https://ijircce.com/admin/main/storage/app/pdf/OUmQRqDgcqyYJ9jHFHGVpo0qIvpQNBV9cNihzyjz.pdf

[45] Pradhan, D. R. (2025) "Generative Agents at Scale: A Practical Guide to Migrating from Dialog Trees to LLM Frameworks," International Journal of Computer Technology and Electronics Communication (IJCTEC) . International Journal of Computer Technology and Electronics Communication (IJCTEC), 8(5), p. 11367. doi: 10.15680/IJCTECE.2025.0805010. https://ijctece.com/index.php/IJCTEC/article/view/230/192

[46] Pradhan, Dr. Rashmiranjan. "Generative Agents at Scale: A Practical Guide to Migrating from Dialog Trees to LLM Frameworks." International Journal of Computer Technology and Electronics Communication (IJCTEC) , vol. 8, no. 5, International Journal of Computer Technology and Electronics Communication (IJCTEC), 2025, p. 11367.

[47] Pradhan, D. R. (2025) "Establishing Comprehensive Guardrails for Digital Virtual Agents: A Holistic Framework for Contextual Understanding, Response Quality, Adaptability, and Secure Engagement," International Journal of Innovative Research in Computer and Communication Engineering.doi:10.15680/IJIRCCE.2025.1307013. https://ijircce.com/admin/main/storage/app/pdf/e9xlTkp5RqODN3RmJOT2uK5biLYlwDggGH9ngoi6.pdf

[48] Pradhan, D. R. (2025). Establishing Comprehensive Guardrails for Digital Virtual Agents: A Holistic Framework for Contextual Understanding, Response Quality, Adaptability, and Secure Engagement. International Journal of

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 8, Issue 5, September – October 2025 ||

DOI: 10.15680/IJCTECE.2025.0805017

Innovative Research in Computer and Communication Engineering. https://doi.org/10.15680/IJIRCCE.2025.1307013

[49] Pradhan DR. Establishing Comprehensive Guardrails for Digital Virtual Agents: A Holistic Framework for Contextual Understanding, Response Quality, Adaptability, and Secure Engagement. International Journal of Innovative Research in Computer and Communication Engineering. 2025; doi:10.15680/IJIRCCE.2025.1307013

[50] Pradhan, Dr. Rashmiranjan. "Establishing Comprehensive Guardrails for Digital Virtual Agents: A Holistic Framework for Contextual Understanding, Response Quality, Adaptability, and Secure Engagement." International Journal of Innovative Research in Computer and Communication Engineering, 2025. doi:10.15680/IJIRCCE.2025.1307013.

[51] Pradhan, D. R. RAGEvalX: An Extended Framework for Measuring Core Accuracy, Context Integrity, Robustness, and Practical Statistics in RAG Pipelines. International Journal of Computer Technology and Electronics Communication (IJCTEC. https://doi.org/10.15680/IJCTECE.2025.0805001

[52] Pradhan, D. R. (2025). RAG vs. Fine-Tuning vs. Prompt Engineering: A Comparative Analysis for Optimizing AI Models. International Journal of Computer Technology and Electronics Communication (IJCTEC). https://doi.org/10.15680/IJCTECE.2025.0805004 https://ijctece.com/index.php/IJCTEC/article/view/170/132