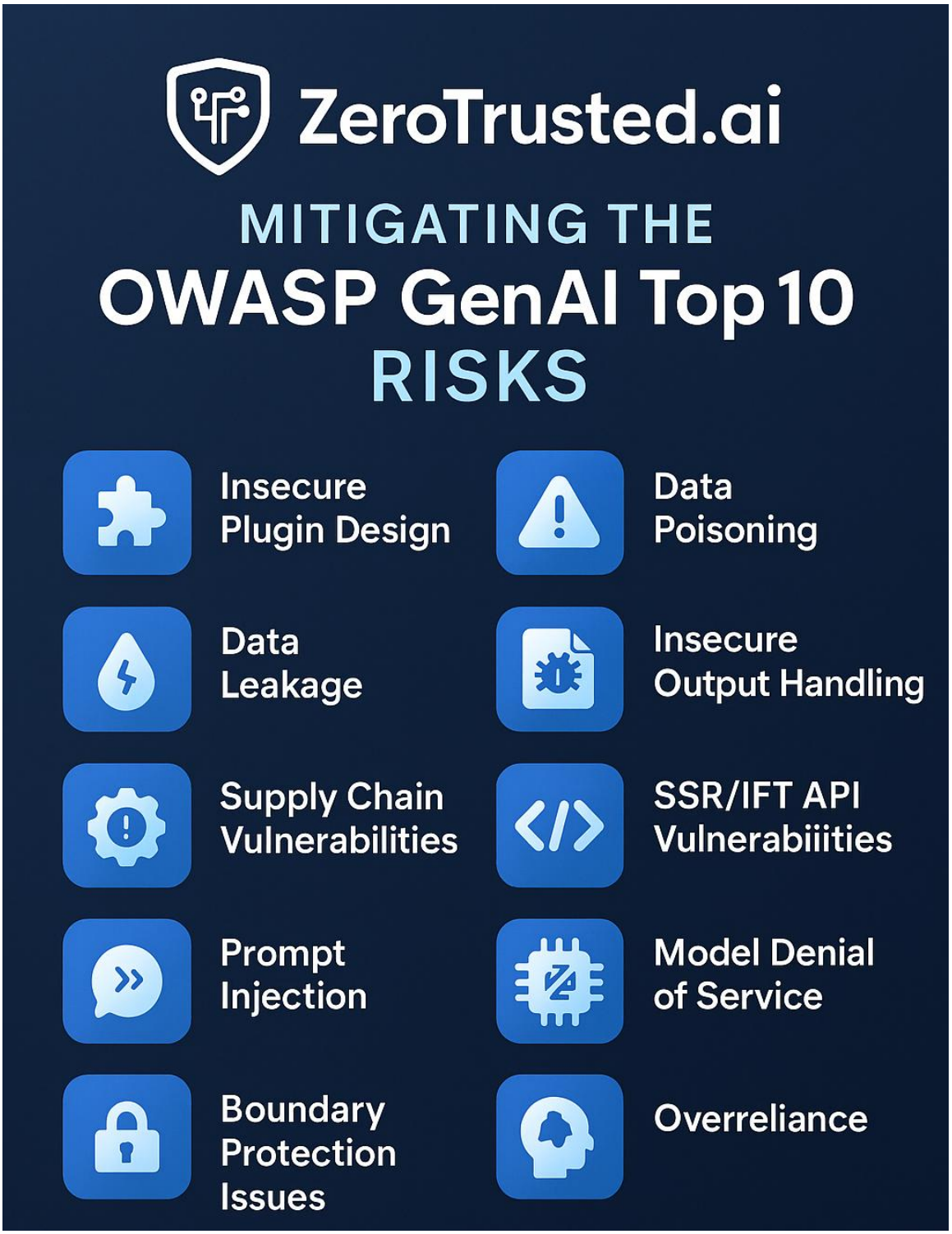
ZeroTrusted.ai
MITIGATING THE
OWASP GenAI Top 10
RISKS
Insecure Plugin Design
Data Poisoning
Data Leakage
Insecure Output Handling
Supply Chain Vulnerabilities
SSR/IFT API Vulnerabiiities
Prompt Injection
Model Denial of Service
Boundary Protection Issues
Overreliance

ZeroTrusted.ai: Mitigating the OWASP GenAI Top 10 Risks

ZeroTrusted.ai provides native Zero Trust security, privacy, and reliability across the full AI lifecycle. Our platform mitigates the most critical vulnerabilities outlined by the OWASP Foundation's **GenAI Top 10**, ensuring AI systems are safe, compliant, and enterprise-ready.

---

1. Prompt Injection

Threat: Attackers manipulate input prompts to alter model behavior.
Mitigation:

- **Secure Prompt Filtering & Validation Engine**
- **AI Firewall** that inspects all prompts pre- and post-processing
- Role-based input validation and natural language sanitization

---

2. Insecure Output Handling

Threat: Malicious or unsafe outputs affect downstream systems or users.
Mitigation:

- **AI Judges** review and score responses before delivery
- Policy enforcement to detect toxicity, PII leakage, and hallucinations – including organizations specific settings
- Explainability and guardrails for high-risk tasks

---

3. Training Data Poisoning

Threat: Adversaries insert malicious data into training sets.
Mitigation:

- **AI HealthCheck** audits training data for poisoning signatures
- Provenance tracing and real-time anomaly detection during training
- Auto-labeling and quarantining of untrusted data

---

4. Model Theft

Threat: Models are stolen or reverse-engineered.
Mitigation:

- Access control to models and agents via **Zero Trust API Gateway**
- Token throttling and watermarking for IP protection
- Usage logging and behavior-based access restrictions

---

5. Model Inversion

Threat: Attackers reconstruct training data from model outputs.
Mitigation:

- Output differential privacy protections
- Response mutation and adaptive noise injection
- Rate-limiting of inference and output depth control

---

6. Sensitive Information Disclosure

Threat: AI leaks PII, IP, or confidential business data.
Mitigation:

- **Content-aware redaction** and **post-query reinjection**
- Data classification engine enforces privacy-aware inference
- Regulatory and sensitive word classification

---

7. Insecure Plugin Ecosystem

Threat: Third-party plugins may introduce unverified risks.
Mitigation:

- **ZeroTrusted.ai Plugin Isolation Sandbox**
- Continuous runtime monitoring and anomaly detection

---

8. Overreliance on AI (Automation Bias)

Threat: Humans blindly trust flawed AI output.
Mitigation:

- Confidence scoring, human-in-the-loop workflows
- Explainable AI (XAI) integration
- Alerts when models lack sufficient evidence or confidence

---

9. Model Misuse or Misalignment

Threat: AI used outside intended purpose or behaves unexpectedly.
Mitigation:

- Enforced policy boundaries based on regulator settings (NIST, PCI, HIPAA) and **alignment testing tools**
- **Usage Monitoring Dashboard** logs all queries by intent
- Drift detection ensures models remain in approved domains

---

10. Supply Chain Vulnerabilities

Threat: Dependencies or third-party models may contain hidden risks.
Mitigation:

- Full **Software Bill of Materials (SBOM)** for AI stacks
- Security scoring and threat assessments for all AI components
- Continuous CVE/SCAP scanning integrated with threat intel

Regulatory Readiness & AI Governance

ZeroTrusted.ai also ensures compliance with:

- NIST AI RMF & SP 800-53 / 100-600 controls
- MIT AI RISK and OWASP AIGen
- OCR Section 1557 (AI Bias & Discrimination)
- HIPAA, GDPR, and CCPA for healthcare and consumer protections

- ISO/IEC 42001 AI Management System Standards (AIMS)

Summary: Why ZeroTrusted.ai?

- **Purpose-built to secure AI, LLMs, Agents, and Vector DBs**
- **Enforces Zero Trust at every AI layer—training, inference, and integration**
- **Backed by 20+ years of cybersecurity expertise, and optimized for mission-critical use cases**