Certified Offensive AI Security Professional (COASP)

EXAM *Blueprint* v1

C|OASP
Certified Offensive AI Security Professional

Certified Offensive AI Security Professional (C|OASP)

Exam Blueprint

S. No.	Domain	Sub-domains	Description/Topics	Weightage	No. of Practical Q.	No. of MCQs.
1.	**Foundations of Offensive AI Security**	Introduction & Evolution of Offensive AI Security	▪ Introduction & Evolution of AI in Security ▪ AI Attack Surfaces ▪ Input Attacks ▪ Enterprise Risk Expansion ▪ Layered AI Threats ▪ Plugin and Extension Risks ▪ AI in Defense, Enterprise, and Government Operations ▪ Security Frameworks for AI Systems ▪ DoD AI Ethical Principles ▪ Data Curation Workflows in AI Security ▪ Assurance Case Methodology ▪ Mission and Business Requirement Mapping ▪ Foundational Knowledge, Skills, and Abilities ▪ Self-assessment & Practice	**5%**	**1**	**2**
		Foundations of LLMs and Agent Architectures	▪ LLM Internals - Prompts, Embeddings, Fine-Tuning, and Context Windows	**6%**		**2**

			▪ LLM Architectural Components ▪ The Inference Pipeline ▪ Fine-tuning and Security Impact ▪ Context Windows and LLM Memory ▪ Agent Architectures - RAG, Plugins, and Tool Invocation ▪ Types of Agent Architectures ▪ Attack and Abuse Scenarios ▪ Attack Surface Identification: Hosted vs. Self-hosted LLMs ▪ Data Curation Workflows - Ensuring Dataset Integrity ▪ Assurance Case Thinking - Proving Trust in LLM Systems ▪ Dataset Assurance KSAs - Metrics & Performance Indicators ▪ DCWF Mapping Summary Table			
2	**AI Reconnaissance and Threat Profiling**	Reconnaissance and Threat Mapping of AI Systems	▪ Enumeration Techniques ▪ Endpoint Fingerprinting ▪ Model Probing ▪ System Prompt Extraction ▪ Reconnaissance Techniques ▪ Identifying Plugins and Context Limits ▪ Case Studies of AI System Leaks ▪ Threat Mapping Workshop ▪ Operational Risk Case Studies	**12%**	**1**	**8**

			▪ DoD-specific Data Policy & Legal Context ▪ Mission and Business Requirement Mapping ▪ DCWF Mapping Summary Table			
		Prompt Injection & Context Exploitation	▪ Prompt Injection Attacks ▪ Context Overflow Exploits ▪ Attack Chain Documentation ▪ Operational Risk Mapping ▪ DoD Data Policy & Compliance Considerations ▪ Mission Mapping for Prompt Exploits	**10%**		**7**
3	**AI System Exploitation Techniques**	Output Exploitation and Memory Manipulation	▪ Trust Boundaries for LLM Output ▪ Output Exploits (XSS, SSRF, RCE) ▪ Agent Memory Attacks ▪ Rare Events and Blind Spot Discovery ▪ Building AI Risk Assessment Frameworks ▪ Quantifying and Communicating AI Risk ▪ Incident Response & Detection Standards	**8%**	**1**	**7**
		Agent Hijacking & Unsafe Tool Integrations	▪ Agent Task Flow Manipulation ▪ Exploiting Excessive Autonomy ▪ Privilege Escalation Attacks ▪ Case Study: Compromised Multi-Agent Orchestration ▪ Human-in-the-Loop Weaknesses ▪ Security Review & Gap Analysis	**8%**		**7**

			▪ Assurance Case Development ▪ Human Factors & UX Testing			
		Supply Chain, Plugin, and Ecosystem Attacks	▪ Malicious Plugin Development ▪ Typosquatting and Version Downgrade Attacks ▪ Shadowing Attacks ▪ Risks from Insecure Integrations ▪ Case Example: Compromised Dependencies ▪ Mission and Business Mapping ▪ AI Solution Integration with Cloud ▪ Ecosystem KSA Mapping	**15%**		**10**
4	**AI Model and Lifecycle Attacks**	Training-time & Lifecycle Attacks	▪ Data Poisoning Campaigns ▪ Targeted vs. Broad Poisoning ▪ Poisoned Embeddings in RAG Pipelines ▪ Adversarial Input Perturbations ▪ Lifecycle Threats in MLOps ▪ Structured Risk Assessment ▪ Trustworthiness Measurement ▪ Operational Realism in Adversarial Testing ▪ MLOps/DevSecOps Integration	**12%**	**1**	**8**
		Model Theft, Extraction, and Evasion Attacks	▪ Query-based Model Extraction ▪ Model Fingerprinting ▪ Jailbreaking and Guardrail Evasion ▪ Building Surrogate Models ▪ Mitigation and Detection Strategies	**8%**		**5**

			▪ AI Verification and Validation Procedures ▪ Resource Estimation for Testing ▪ Automation Workflows ▪ New KSAs for Extraction & Incident Response			
5	**Governance and Defensive Engineering**	Governance, Risk, & Compliance	▪ AI-specific Test & Evaluation Planning ▪ Go/No-Go Decision Frameworks ▪ Linking Assurance Cases to Compliance ▪ Safety Standards (MIL-STD 882E, DO-178C, ISO 26262) ▪ Compliance KSAs ▪ Professional Report Writing ▪ Compliance Mapping	**8%**	**1**	**5**
		Defensive Engineering & Mitigation Strategies	▪ Input/Output Filtering ▪ Sandboxing and Least Privilege ▪ Monitoring AI Behavior Drift ▪ Blue Team Responses for AI Incidents ▪ Case Study: Securing RAG Pipelines ▪ Bias & Representation Testing ▪ Framework Integration ▪ ITIL Alignment	**8%**		**4**
		Total		**100%**	**5**	**65**