Module 1 – Terminology

TERM	DEFINITION
Core Red Teaming Concepts	
Adversarial Thinking	Mindset of identifying and exploiting weaknesses like an attacker.
Attack Surface	All possible entry points a system exposes to potential attackers.
Blue Team	Defensive group responsible for protecting systems.
Engagement	A formal red team operation or campaign.
Penetration Testing (Pentest)	Controlled test of systems for vulnerabilities; narrower in scope than red teaming.
Purple Team	Collaborative effort between red and blue teams to improve defenses.
Red Team	Group simulating adversaries to identify weaknesses.
Rules of Engagement (RoE)	Predefined scope, limits, and conditions for a red team exercise.
Threat Model	Structured analysis of adversary goals, capabilities, and possible attack paths.
AI & Frontier Model Red Teaming	
Adversarial Example	Intentionally crafted input that causes an AI to misclassify or fail.
Alignment Red Teaming	Testing whether an AI stays within its intended ethical or safety guidelines.
Backdoor Attack	Embedding hidden triggers in training data to activate malicious behavior later.
Boundary Testing	Pushing inputs to edge cases where a model might fail.
Capability Red Teaming	Testing the full range of what an AI can do, including harmful capabilities.
Data Poisoning	Injecting malicious data into training / fine-tuning datasets.
Evasion Attack	Crafting inputs that avoid model detection (e.g., bypassing content filters).
Frontier Model	The most advanced AI systems available at a given time.
Hallucination	Confident but factually incorrect AI output.
Indirect Prompt Injection	Prompt injection hidden in external sources (e.g., websites, docs) that a model processes.
Jailbreak	Input that bypasses an AI model's safeguards.
Membership Inference Attack	Determining whether specific data was part of a model's training set.

TERM	DEFINITION
Mode Collapse	A generative model outputting repetitive or overly narrow results due to degraded diversity.
Model Inversion	Extracting sensitive training data from a model.
Model Stealing	Reconstructing a model’s functionality by querying it extensively.
Multimodal Attack	Exploiting models across modalities (text + image + audio).
Over-Optimization Exploit	Driving a model to optimize too aggressively for a metric, causing harmful side effects.
Prompt Chaining	Combining multiple prompts or models to induce unexpected behaviors.
Prompt Injection	Malicious prompt that manipulates a model into unintended behavior.
Safety Red Teaming	Focused evaluation of whether an AI system avoids disallowed, unethical, or unsafe behaviors.
Social & Psychological Red Teaming	
Baiting	Luring a victim with something attractive (USB, download, etc.).
Deepfake Attack	Using AI-generated media to impersonate a person for manipulation.
Impersonation	Pretending to be a trusted person or entity.
Pretexting	Creating a fabricated scenario to manipulate a victim.
Psychological Exploit	Leveraging biases, trust, fear, or urgency to influence behavior.
Social Engineering	Manipulating people into revealing information or taking harmful actions.
Tailgating / Piggybacking	Physically following someone into a restricted area.
Reporting & Documentation	
After-Action Report (AAR)	Comprehensive document summarizing findings of an engagement.
Executive Summary	Non-technical overview of risks for leadership.
Lessons Learned	Insights for improving future defenses or testing campaigns.
Mitigation	Steps to reduce or eliminate identified risks.
Remediation	Fixing the underlying vulnerabilities discovered.
Risk Assessment	Evaluation of severity and likelihood of discovered vulnerabilities.
Severity Rating	Classification of vulnerabilities (e.g., Critical, High, Medium, Low).

TERM	DEFINITION
Technical Appendix	Detailed exploit walkthroughs, payloads, logs, evidence.
Team & Operational Terms	
Analyst	Member focused on research, intelligence, and documentation.
Black Team	Aggressor group with minimal disclosure; often operates covertly.
Campaign	Multi-phase, long-term red teaming operation.
Engagement Lifecycle	Plan → Execute → Report → Review.
Operator	Individual executing technical aspects of attacks.
Red Cell	A specialized red team unit, often in military/intelligence contexts.
White Team	Oversees exercises, ensures safety, adjudicates results.
Writer / Reporter	Member responsible for producing engagement reports.
Evaluation & Testing	
Capture the Flag (CTF)	Competitive exercise simulating real-world adversarial problems.
Edge Case Testing	Exploring rare or extreme conditions that cause failure.
False Negative	Failing to detect actual malicious activity.
False Positive	Incorrectly identifying benign behavior as malicious.
Resilience Testing	Evaluating how well a system withstands sustained attack.
Scenario Simulation	Running hypothetical adversarial situations against a system.
Stress Testing	Pushing a system beyond normal limits to expose weaknesses.