AI and LLM Security Questions & Answers

Question	Classic AI Answer	GenAI Answer
What is the difference between adversarial attacks on traditional ML models versus LLMs?	Adversarial attacks on traditional ML models typically involve small perturbations to input data (like adding noise to images) that cause misclassification while remaining imperceptible to humans. Examples include FGSM and PGD attacks on image classifiers.	LLM adversarial attacks focus on prompt manipulation and jailbreaking techniques. Attackers craft specific prompts to bypass safety guardrails, extract training data, or generate harmful content. Unlike traditional ML, these attacks exploit the model's language understanding and generation capabilities through carefully constructed text inputs.
What are the main types of data poisoning attacks in AI systems?	Data poisoning involves injecting malicious samples into training data to degrade model performance or create backdoors. Types include targeted attacks (affecting specific inputs) and untargeted attacks (general performance degradation). Examples include label flipping and backdoor attacks with trigger patterns.	In LLMs, data poisoning can occur through web scraping contaminated content, instruction tuning with malicious examples, or RLHF manipulation. Attackers may inject biased content, misinformation, or backdoor triggers into training corpora. The scale and diversity of LLM training data make detection particularly challenging.
How do model extraction attacks work?	Model extraction attacks involve querying a target model to steal its functionality or parameters. Attackers send crafted inputs and analyze outputs to reconstruct the model architecture, weights, or decision boundaries. This threatens IP protection and enables further attacks.	LLM extraction attacks focus on recreating model capabilities through prompt engineering and output analysis. Attackers may use techniques like few-shot learning prompts, API querying patterns, or knowledge distillation to build surrogate models. The challenge is extracting complex reasoning and generation capabilities rather than simple classification boundaries.
What is prompt injection and how does it differ from traditional input validation attacks?	Traditional input validation attacks like SQL injection exploit parsing vulnerabilities in structured systems. Attackers inject malicious code into input fields that gets executed by the backend system, often targeting databases or web applications with predictable parsing logic.	Prompt injection exploits the natural language processing capabilities of LLMs. Attackers embed malicious instructions within seemingly normal prompts to manipulate the model's behavior. Unlike traditional injection attacks, these leverage the model's understanding of context and instructions rather than exploiting parsing vulnerabilities.
What are the privacy risks specific to large	Traditional ML privacy risks include membership inference attacks	LLMs face unique privacy challenges including training data memorization and regurgitation,

Question	Classic AI Answer	GenAI Answer
language models?	(determining if data was in training set), model inversion attacks (reconstructing training data), and attribute inference attacks (learning sensitive attributes). Differential privacy and federated learning are common mitigations.	where models can reproduce verbatim text from training data. Risks include exposure of PII, copyrighted content, and sensitive information through careful prompting. The vast scale of training data makes comprehensive privacy auditing extremely difficult.
How do alignment attacks target AI safety measures?	Traditional AI safety focuses on robustness, reliability, and preventing unintended behaviors. Attacks typically target specific failure modes or edge cases in well-defined tasks. Safety measures include adversarial training, formal verification, and robust optimization techniques.	Alignment attacks specifically target the human preference alignment of LLMs. Techniques include jailbreaking prompts, role-playing scenarios, and multi-turn conversations that gradually shift the model away from its safety training. Attackers exploit the tension between helpfulness and harmlessness in instruction-following models.
What are the challenges in detecting AI-generated content for security purposes?	Traditional AI detection focuses on identifying deepfakes, manipulated images, or synthetic media. Detection methods include statistical analysis, metadata examination, and specialized neural networks trained to identify artifacts from generation processes.	Detecting LLM-generated text presents unique challenges due to the high quality and diversity of outputs. Detection methods include perplexity analysis, stylometric features, and watermarking techniques. However, the rapid improvement in generation quality and the ability to fine-tune detection-resistant models make this an ongoing arms race.
How do supply chain attacks affect AI/ML systems?	Traditional ML supply chain attacks target model repositories, training pipelines, or deployment infrastructure. Attackers may compromise datasets, inject malicious code into ML libraries, or tamper with model files during distribution.	LLM supply chain attacks can target foundation model providers, fine-tuning services, or plugin ecosystems. Risks include compromised pre-trained models, malicious fine-tuning datasets, or backdoored API integrations. The complexity of LLM deployment stacks creates multiple attack surfaces across the supply chain.
What is the difference between model robustness and model security?	Model robustness focuses on maintaining performance under natural distribution shifts, noise, or edge cases. It's about reliability and consistent behavior across	In LLMs, robustness involves consistent performance across diverse prompts and contexts, while security addresses intentional attempts to manipulate model behavior.

Question	Classic AI Answer	GenAI Answer
	different conditions. Security specifically addresses malicious attacks and adversarial scenarios designed to exploit vulnerabilities.	Security encompasses prompt injection resistance, alignment maintenance, and preventing harmful output generation, extending beyond traditional robustness concerns.
How do multi-modal AI systems introduce new security vulnerabilities?	Traditional multi-modal systems combining different data types (text, images, audio) face challenges in unified processing and decision-making. Vulnerabilities often arise from inconsistencies between modalities or exploitation of the fusion process.	Multi-modal LLMs face unique security challenges including cross-modal prompt injection (using images to inject text instructions), modality-specific jailbreaking, and attacks that exploit the model's reasoning across different input types. The complexity of processing and aligning multiple modalities creates new attack surfaces not present in single-modal systems.