# Generated Quiz

## Multiple Choice Questions

1. 1. What defines a 'Prompt Injection Vulnerability'?

* A) A system's inability to process multimodal inputs.
* B) A security flaw where user prompts alter an LLM’s behavior or output in unintended ways.
* C) The accidental revelation of an LLM's internal system prompts.
* D) An LLM generating false or inaccurate information.

Correct Answer: A security flaw where user prompts alter an LLM’s behavior or output in unintended ways.

1. 2. Which of the following best describes 'Jailbreaking' in the context of LLMs?

* A) The process of an LLM processing external content with malicious instructions.
* B) An LLM revealing sensitive information to a user.
* C) A specific form of prompt injection causing an LLM to completely disregard its safety protocols.
* D) The manipulation of training data to introduce biases.

Correct Answer: A specific form of prompt injection causing an LLM to completely disregard its safety protocols.

1. 3. What is characteristic of an 'Indirect Prompt Injection'?

* A) A user's directly provided input prompt intentionally alters the LLM's behavior.
* B) An LLM processes external content containing malicious instructions that alter its behavior.
* C) The LLM reveals its underlying system prompts.
* D) The LLM generates excessive output without constraint.

Correct Answer: An LLM processes external content containing malicious instructions that alter its behavior.

1. 4. The 'RAG Triad' is a set of criteria used to evaluate LLM responses. What three aspects does it focus on?

* A) Processing speed, data privacy, and ethical considerations.
* B) Context relevance, groundedness (accuracy relative to source material), and question/answer relevance.
* C) Input validation, output sanitization, and resource consumption.
* D) Model size, training data volume, and deployment cost.

Correct Answer: Context relevance, groundedness (accuracy relative to source material), and question/answer relevance.

1. 5. According to LLM02:2025, what vulnerability is defined as 'Sensitive Information Disclosure'?

* A) An attacker manipulating training data to introduce backdoors.
* B) An LLM or its application context revealing confidential data like PII or financial details.
* C) An LLM generating false or misleading information.
* D) An LLM being granted excessive autonomy.

Correct Answer: An LLM or its application context revealing confidential data like PII or financial details.

1. 6. What does 'Data and Model Poisoning' (LLM04:2025) refer to?

* A) The accidental revelation of an LLM's system prompts.
* B) An LLM generating outputs that are not properly validated or sanitized.
* C) Malicious manipulation of data used in pre-training, fine-tuning, or embeddings to introduce vulnerabilities or biases.
* D) The compromise of training data integrity within the LLM's development ecosystem.

Correct Answer: Malicious manipulation of data used in pre-training, fine-tuning, or embeddings to introduce vulnerabilities or biases.

1. 7. The vulnerability 'Excessive Agency' (LLM06:2025) describes what condition in an LLM-based system?

* A) The LLM producing false, inaccurate, or misleading information.
* B) The system is granted a greater degree of autonomy or ability to call functions than is necessary or safe.
* C) The LLM revealing confidential data such as PII.
* D) The LLM generating outputs based on inputs without adequate constraints.

Correct Answer: The system is granted a greater degree of autonomy or ability to call functions than is necessary or safe.

## True/False Questions

1. 1. A Prompt Injection Vulnerability occurs when user prompts alter the LLM’s behavior or output in unintended ways.

Answer: True

1. 2. Techniques like Retrieval Augmented Generation (RAG) and fine-tuning fully mitigate all prompt injection vulnerabilities.

Answer: False

1. 3. Jailbreaking is a form of prompt injection.

Answer: True

1. 4. Indirect prompt injections occur when an LLM accepts input from external sources that alters its behavior.

Answer: True

1. 5. Multimodal AI introduces unique prompt injection risks.

Answer: True

1. 6. MITRE ATLAS AML.T0051.000 refers to LLM Prompt Injection: Indirect.

Answer: False

1. 7. Implementing input and output filtering is a listed mitigation strategy for prompt injection.

Answer: True