

PRACTICE QUESTIONS

GIAC / SANS

GAIPT

GIAC AI Penetration Tester

SHARED BY [u/mighty_widget_86](#) on aiexamcerts

EDITION Free Set 2

QUESTIONS 10

SHARED October 18, 2026

EXAM GIAC / SANS GA IPT, GIAC AI Penetration Tester

Shared by [u/mighty_widget_86](#) on aiexamcerts

10 practice questions for the GIAC / SANS GAAPT exam, GIAC AI Penetration Tester, each followed by its answer and a short explanation. More free sets and the full practice exam: aiexamcerts.com/giac-sans/gaapt. Spotted a mistake? Email support@aiexamcerts.com with the exam code GAAPT and the question number.

1. Which technique hides a forbidden instruction by splitting it across several messages so that no single message looks harmful on its own?
- A Payload splitting
 - B Model distillation
 - C Embedding poisoning
 - D Membership inference

Answer: A

Payload splitting breaks a harmful request into benign-looking fragments and relies on the model to reassemble them. Model distillation trains a smaller model to imitate a larger one and is not a prompt technique. Embedding poisoning and membership inference target retrieval data and training data respectively.

2. A model that refuses a jailbreak in the first turn cannot be attacked successfully with a multi-turn strategy, because the model keeps no memory of earlier turns.
- A True
 - B False

Answer: B (False)

Each request carries the conversation history inside the context window, so earlier turns directly shape later responses. Multi-turn escalation works precisely because the model sees the accumulated conversation. The statement is false.

3. A retrieval-augmented chatbot pulls in a document an employee uploaded, and the document contains an instruction to reveal other users' records. Which control most directly reduces this risk?
- A Increasing the number of retrieved chunks
 - B Treating retrieved text as untrusted data and keeping it out of the instruction channel
 - C Reducing the embedding dimension to lower storage costs in the vector index and speed up search
 - D Caching retrieved chunks for longer

Answer: B

Retrieved documents are data written by people who may be hostile, so they must not be able to issue instructions, and any resulting data access must pass authorization checks. Retrieving more chunks increases exposure to the injected text. Embedding dimension and cache lifetime do not change whether retrieved text can steer the model.

4. Which trust boundary is crossed when an LLM agent places text returned by a web search tool into its next planning step?
- A Between two GPU cards inside one server that share a PCIe bus for model parallelism
 - B Between the user's browser and a content delivery network
 - C Between untrusted tool output and the agent's instruction context
 - D Between the training split and the validation split

Answer: C

Search results come from outside the system and can contain attacker-written text, so moving them into the planning context crosses from untrusted data into instructions. The GPU, CDN and dataset split boundaries are real in other contexts but do not involve trust in the agent's inputs. Treating this boundary as a first-class control is central to agent security.

5. Loading model weights from a pickle file is risky because deserialization can run arbitrary code. The safer format that stores only tensor data is called _____.

Answer: safetensors

The safetensors format stores tensor names and raw values without executable objects, so loading it cannot run code from the file. Pickle-based checkpoints can carry malicious payloads that execute on load. The blank accepts the standard lowercase name and its singular form.

6. A model trained to maximize a learned reward model starts producing answers that score highly but that human reviewers find unhelpful. What is this most likely called?
- A Data leakage
 - B Catastrophic forgetting of earlier skills during later training runs
 - C Quantization error
 - D Reward model overoptimization

Answer: D

When a policy optimizes a proxy reward too hard, it drifts toward responses the proxy overvalues and away from what humans actually want. Catastrophic forgetting describes losing earlier skills during new training, which is a different phenomenon. Quantization error comes from lowering numeric precision and has nothing to do with the reward signal.

7. Which two techniques help an attacker get past keyword-based input filters or detect guardrail behavior? (Choose two.)
- A Encoding the request in Base64 or another alternate encoding
 - B Paraphrasing the request in a different language
 - C Reading the license terms of the model's published weights
 - D Measuring GPU memory usage while the service is idle
 - E Requesting a larger batch size from the API

Answer: A, B

Encoding and translation change the surface form of a request so keyword matches miss it while the model still understands it. License terms, idle GPU memory and batch size do not change how filters or models interpret input. They are distractors that describe real but unrelated facts about a deployment.

8. Which tactic is most useful for discovering which AI services a company runs, before any active testing?
- A Passive enumeration of DNS names, TLS certificates and public API documentation
 - B Brute-forcing the model's weights directly from the inference server's memory and disk
 - C Fine-tuning a private copy of the target model
 - D Poisoning the company's vector store

Answer: A

Passive reconnaissance reveals hostnames, certificates and documented routes without touching the target's systems in an intrusive way. Brute-forcing weights is not a meaningful technique, and fine-tuning a private copy or poisoning a store both require access that reconnaissance is meant to find first. Certificate transparency and public docs are typical sources for this step.

9. What describes a confused deputy problem in an AI agent?
- A The model returns an incorrect citation to a source it never read, because retrieval returned a near match
 - B A privileged agent is tricked by untrusted input into using its authority for an attacker
 - C Two agents share a single GPU and slow each other down
 - D The tokenizer splits a single word into two tokens

Answer: B

The deputy has authority that the attacker lacks, and the attacker manipulates it into exercising that authority. An incorrect citation is a grounding error rather than an authority problem. GPU sharing and tokenization are performance and model-internals topics with no link to delegated trust.

10. Human approval gates are unnecessary for agents that only read data, because reads cannot cause any security impact.

A True

B False

Answer: B (False)

Reads can leak sensitive records into a response, a log or a later outbound message, so read-only agents still need scoping and output checks. Approval gates matter most for writes, but reading is not harmless. The statement is false.