PUBLICATION

EXAM-TM40H-PRE

PRE-TEST — SL 4H: AI ENGINEER

Maven Smart System (MSS) — USAREUR-AF

HEADQUARTERS
UNITED STATES ARMY EUROPE AND AFRICA
(USAREUR-AF)
Wiesbaden, Germany

26 MARCH 2026

DRAFT — UNOFFICIAL — NOT FOR OPERATIONAL USE

PRE-TEST — SL 4H: AI ENGINEER

MAVEN SMART SYSTEM (MSS) — USAREUR-AF

Field	Detail
Course	SL 4H: AI Engineer
Form	Pre-Test
Level	SL 4H (Specialist)
Audience	AI/ML specialists; prerequisite: SL 1+20+30 + Python + prompt engineering familiarity
Time Allowed	30 minutes
Passing Score	N/A — diagnostic only

INSTRUCTIONS

This diagnostic assessment establishes your baseline knowledge before training. Your score does not affect course eligibility. Answer honestly — results help the instructor tailor instruction to gaps.

SECTION 1 — MULTIPLE CHOICE

Circle the letter of the best answer. (2 points each)

1. A large language model (LLM) generates text by:

A. Retrieving stored answers from a database of pre-approved responses B. Executing deterministic rule-based logic against a structured knowledge base C. Predicting the most likely next token(s) in a sequence based on learned probability distributions from training data D. Running a nearest-neighbor search against vector embeddings of the input query

2. "Prompt engineering" refers to:

A. Writing Python code that calls an LLM API B. Fine-tuning an LLM on domain-specific training data C. Configuring the inference endpoint for an LLM deployment D. Crafting and structuring input text to guide an LLM toward desired outputs

3. In AI system design, "human-in-the-loop" means:

A. A human monitors system uptime but does not review individual outputs B. Human feedback is used to fine-tune the model after deployment C. A human reviews and approves AI-generated outputs before they are acted upon or distributed D. The AI system notifies a human when it encounters low-confidence outputs

4. "Hallucination" in the context of LLMs refers to:

A. Outputs that are intentionally misleading due to adversarial inputs B. An output that cites a real source but misquotes it C. A model that produces inconsistent outputs across repeated identical prompts D. The model generating plausible-sounding but factually incorrect or fabricated information

5. "Retrieval-Augmented Generation" (RAG) improves LLM output quality by:

A. Retrieving relevant documents at inference time and including them in the prompt as context B. Fine-tuning the LLM on domain-specific data before deployment C. Increasing the size of the LLM's context window to include all available documents D. Running multiple LLM instances in parallel and aggregating their outputs

6. The "context window" of an LLM limits:

A. The total number of tokens (input + output) the model can process in a single inference call B. The number of API calls that can be made per minute C. The number of documents that can be stored in the retrieval corpus D. The maximum number of users who can query the model simultaneously

7. Which of the following is an example of an appropriate use of AI in an Army operational context?

A. Autonomous lethal targeting without human review B. AI-generated operations orders published directly to the field without commander review C. Automated release of SITREP data to coalition partners without classification review D. AI-generated draft summary of maintenance trends for analyst review and approval before distribution

8. "Structured output" from an LLM (e.g., JSON format) is useful in automated pipelines because:

A. JSON outputs are automatically encrypted for secure transmission B. Structured outputs can be parsed programmatically by downstream systems without additional processing C. The LLM produces more accurate content when forced to use JSON format D. Structured outputs bypass the context window limitation

9. "Chain-of-thought" prompting is a technique where:

A. The prompt instructs the model to reason step-by-step before producing a final answer B. Multiple LLM calls are made sequentially with each output feeding the next input C. The prompt includes example input-output pairs to guide the model D. The model's output is validated by a second LLM call

10. An AI system that autonomously submits data to an operational system without any human review step would violate which principle?

A. Least-privilege access control B. Data minimization policy C. Human-in-the-loop (HITL) requirement D. Classification handling procedures

11. The Army CIO Memorandum (April 2024) on AI use establishes which of the following requirements?

A. All Army AI systems must be operated only by FA49 officers B. Generative AI tools may not be used for any Army administrative functions C. AI-generated content must be reviewed by a human before official use, and use cases must comply with applicable policies D. Only Government-owned AI models may be deployed on Army networks

12. "Prompt injection" is an adversarial technique where:

A. An attacker modifies the LLM's model weights during inference B. Malicious instructions embedded in user input or retrieved documents attempt to override the system's intended behavior C. An attacker floods the LLM inference endpoint with requests to cause denial of service D. A user extracts the system prompt through repeated queries

13. In a pipeline that uses an LLM to summarize incoming intelligence reports, "chunking" the reports before sending to the LLM is necessary because:

A. Smaller chunks are encrypted more efficiently B. Reports may exceed the LLM's context window limit; chunking breaks them into segments that fit within the limit C. The LLM produces higher-quality summaries when given shorter inputs regardless of length D. Chunking is required by Army policy for all AI document processing

14. "Few-shot prompting" uses:

A. A minimal number of parameters in the prompt to reduce token cost B. Example input-output pairs included in the prompt to demonstrate the desired response format C. A reduced context window to improve inference speed D. A subset of the training corpus for fine-tuning the model

15. Which of the following is a key risk of deploying an LLM-based system into an operational Army context without red-teaming?

A. The system may produce outputs that are incorrect, biased, or manipulable in ways that were not anticipated during development B. The model will require more compute resources than anticipated C. The deployment will violate software licensing agreements D. Users will overload the inference endpoint during peak operational periods

SECTION 2 — SHORT ANSWER

Answer in 2–5 sentences. (6 points each)

SA-1. Explain the difference between fine-tuning an LLM and using RAG. In what scenario would you prefer RAG over fine-tuning for an Army operational use case, and why?

SA-2. A colleague proposes building an AI workflow that automatically publishes AI-generated SITREP summaries to the unit SharePoint without any review step, arguing it will "save time." What is your response, and what procedure should be followed instead?

SA-3. Describe what a "system prompt" is in an LLM deployment and explain why it must be protected from disclosure to end users in an operational AI system.

SA-4. You are asked to evaluate whether an AI use case is appropriate for the Army environment. List four questions you would ask to assess the use case, and explain why each matters.

SA-5. Explain what "token" means in the context of LLMs and why token limits matter for operational AI workflows that process long documents such as orders, AARs, or maintenance logs.

SCORING SUMMARY

Section	Questions	Points Each	Total Points
Multiple Choice	15	2	30
Short Answer	5	6	30
Total	—	—	**60**

Passing: N/A — Pre-test is diagnostic only.

ANSWER KEY — INSTRUCTOR USE ONLY

Do not distribute to students.

Multiple Choice: 1. C — LLMs generate text by predicting next tokens from learned probability distributions. 2. D — Prompt engineering = crafting input text to guide LLM outputs. 3. C — HITL = human reviews and approves outputs before action or distribution. 4. D — Hallucination = plausible but factually incorrect or fabricated content. 5. A — RAG retrieves documents at inference time to include as context. 6. A — Context window limits total tokens (input + output) per inference call. 7. D — AI-generated draft for human review before distribution is appropriate. 8. B — Structured outputs can be parsed programmatically by downstream systems. 9. A — Chain-of-thought = prompt instructs model to reason step-by-step. 10. C — Autonomous action without human review violates the HITL requirement. 11. C — Army CIO Memo (April 2024): AI content requires human review; use cases must comply with policy. 12. B — Prompt injection = malicious instructions attempting to override intended behavior. 13. B — Chunking breaks documents into context-window-sized segments. 14. B — Few-shot prompting uses example pairs in the prompt to demonstrate desired format. 15. A — Without red-teaming, incorrect, biased, or adversarially manipulable outputs may reach operational use.

Short Answer Guidance:

SA-1. Full credit: fine-tuning modifies the model's weights on domain data — expensive, slow, and requires retraining when data changes; RAG retrieves current documents at inference time without changing the model — better for operational data that changes frequently; Army preference for RAG in

most operational cases: classified documents can't be shared for fine-tuning, and operational situation data is constantly updated. Partial credit (3 pts) for correct distinction without Army-specific reasoning.

SA-2. Full credit: reject the approach — automated publishing without review violates the human-in-the-loop requirement and Army CIO policy; the correct procedure is: AI generates a draft SITREP summary → authorized human reviewer reviews for accuracy and completeness → reviewer approves and publishes. "Saving time" does not override HITL requirements, especially for official operational documents. Full credit requires citing HITL requirement and correct procedure.

SA-3. Full credit: a system prompt is a hidden instruction set provided to the LLM before the user's input — it defines the AI's role, constraints, and behavior; it must be protected because disclosure allows users to understand and bypass the constraints, craft adversarial prompts to exploit system behavior, or extract sensitive operational configuration details embedded in the prompt. Partial credit (3 pts) for correct definition without protection rationale.

SA-4. Full credit: any four from — Is there a human review step before any operational action is taken? Does the AI handle classified data, and at what classification level will it operate? Is the use case on the prohibited list (lethal autonomous targeting, personnel evaluation without human review, etc.)? Who is accountable if the AI output is wrong? Is the AI output reversible if it is wrong? Has the use case been reviewed against Army CIO and command AI policy? Has the system been red-teamed? Each question must include a brief "why it matters" statement.

SA-5. Full credit: a token is roughly a word piece — LLMs process text as sequences of tokens (approx. 4 chars per token); context window limits the total tokens in + out per call; long operational documents (orders, AARs, maintenance logs) may exceed the context window — solutions include chunking, summarization, or RAG to retrieve only relevant sections; the limit matters operationally because exceeding it either truncates the document (losing information) or causes the API call to fail. Partial credit (3 pts) for correct definition without operational implication.

USAREUR-AF Operational Data Team TM-40H Pre-Test | Version 1.0 | March 2026