PUBLICATION

EXAM-TM50H-PRE

PRE-TEST — SL 5H: ADVANCED 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 5H: ADVANCED AI ENGINEER

MAVEN SMART SYSTEM (MSS) — USAREUR-AF

Field	Detail
Course	SL 5H: Advanced AI Engineer
Form	Pre-Test
Level	SL 5H (Advanced Specialist)
Audience	Senior AI engineers; prerequisite: SL 4H + production AI experience
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. In a multi-agent AI system, a "circuit breaker" pattern is used to:

A. Limit the compute cost of running multiple agents concurrently B. Monitor agent communication for adversarial content injection C. Route agent tasks to the appropriate specialized sub-agent D. Stop an agent chain from continuing to execute when a failure or anomalous condition is detected, preventing cascading failures

2. "Shared state" in a multi-agent system creates which primary risk?

A. Multiple agents reading and writing shared state can produce race conditions and inconsistent outputs if not managed with proper coordination mechanisms B. State synchronization overhead reduces inference speed C. Shared state is prohibited on Army networks due to security policy D. Agents with shared state cannot be individually monitored or audited

3. LLM "fine-tuning" on an operational corpus is fundamentally different from RAG because:

A. Fine-tuning is faster to deploy than RAG B. Fine-tuning modifies the model's weights, encoding domain knowledge into the model itself — it cannot be quickly updated without retraining C. RAG requires more compute than fine-tuning at inference time D. Fine-tuning produces more deterministic outputs than RAG

4. Fine-tuning an LLM on Army operational corpora requires SJA review because:

A. LLM fine-tuning is classified as a weapons system modification under Army Regulation B. The review determines whether the fine-tuned model qualifies as a new system requiring an ATO C. SJA review is required for all ML model training on government networks D. Operational corpora may contain sensitive information, and encoding it into a model's weights creates persistent knowledge that may be accessible in unauthorized contexts

5. "Hybrid retrieval" in an advanced RAG architecture combines:

A. Two separate LLM inference calls — one for retrieval and one for generation B. Retrieval from classified and unclassified corpora simultaneously C. Dense (semantic embedding) and sparse (keyword/BM25) retrieval signals to capture both semantic similarity and exact-match relevance D. RAG and fine-tuning in the same inference pipeline

6. "Re-ranking" in a RAG pipeline refers to:

A. Re-ordering documents in the corpus by recency before retrieval B. Re-running the embedding model to update document vectors C. A post-retrieval step that reorders candidate documents by relevance quality before injecting them into the prompt context D. The process of removing low-quality documents from the corpus

7. Adversarial prompt injection testing in SL 5H must be conducted:

A. In the production environment to test real-world resilience B. Only by red team personnel with SECRET clearance C. In an isolated test environment — never in production where injected content could affect real data or operations D. After deployment, using production logs to identify real injection attempts

8. "AI observability" in a production AI system includes monitoring:

A. User login counts and session durations B. Model weight updates applied during online learning C. Output quality metrics, input distribution shifts, latency trends, and anomalous output patterns that may indicate model degradation or adversarial activity D. Compute and memory utilization on the inference server

9. The DoD Responsible AI Implementation Taskforce (RAIMTF) guidelines (2024) require which of the following for operational AI systems?

A. All DoD AI systems must be approved by the Joint AI Center before deployment B. AI output review periods of 90 days are required before any operational use C. All operational AI systems must use only Government-developed models D. AI systems must be designed with responsible AI principles: explainability, reliability, governability, traceability, and bias minimization

10. The Army CIO Memorandum (April 2024) on generative AI requires which specific human-in-the-loop provision?

A. AI-generated content must be reviewed by a human before it is used for official purposes, and the human reviewer bears responsibility for the content B. All AI outputs must be reviewed by a GO before official release C. AI systems may not generate text longer than 500 words without human segmentation review D. AI-generated content must be watermarked to identify it as machine-generated

11. "Rate-limiting" for an agent chain that can write to the Foundry Ontology is important because:

A. The Foundry API enforces rate limits that will cause failures if exceeded B. Uncontrolled agent chains could execute a large number of irreversible writes before a human can intervene, requiring careful rate controls to bound the blast radius of any failure C. Rate limiting improves agent inference speed by reducing API call overhead D. Army policy limits all automated writes to a maximum of 100 records per hour

12. In enterprise AI governance, "classification-level constraint for inference endpoints" means:

A. The AI model's security classification determines what network it can be accessed from B. An inference endpoint configured for a specific classification level must only process data at or below that level — cross-level inference is prohibited C. Classification labels must be present in all prompts submitted to the inference endpoint D. AI models used for classified data must be separately trained from unclassified models

13. An AI system that produces outputs that drift gradually over time without any model changes is most likely exhibiting:

A. Input distribution shift — the statistical properties of incoming data have changed, causing the model's responses to drift even though model weights are unchanged B. Catastrophic forgetting from online fine-tuning C. Prompt injection attacks from adversarial users D. Output degradation due to inference endpoint load

14. "Failure isolation" in a multi-agent system is achieved by:

A. Designing agents so that a failure in one agent does not propagate to other agents — using circuit breakers, timeouts, and independent state management B. Running all agents on separate compute instances C. Monitoring each agent's output independently without shared logging D. Using a separate LLM model for each agent to prevent model-level correlation

15. The most significant architectural risk of a multi-agent system where agents share a single writable Ontology namespace is:

A. Performance degradation from concurrent read operations B. A buggy or compromised agent writing incorrect data could corrupt the shared namespace, affecting the inputs and outputs of all other agents in the system C. Classification boundary violations between agent outputs D. Inability to audit which agent produced which output record

---

SECTION 2 — SHORT ANSWER

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

SA-1. Explain the difference between RAG and fine-tuning for incorporating Army-specific terminology and procedures into an LLM system. For each approach, describe one scenario where it would be the better choice and one significant risk.

SA-2. Describe the red-teaming methodology you would apply to a deployed AIP Logic workflow that processes incoming intelligence summaries. What specific adversarial scenarios would you test, and how would you document the results?

SA-3. A multi-agent system is designed to: (1) retrieve relevant doctrine documents, (2) generate a draft operational summary, and (3) route the summary for human review. Describe the circuit breaker and rate-limiting controls you would implement for this architecture.

SA-4. An AI engineering team proposes fine-tuning an LLM on the USAREUR-AF operational lessons-learned corpus. Walk through the required review and approval process before this fine-tuning can proceed, including the SJA review and classification considerations.

SA-5. Describe what "AI observability" means at the enterprise level for a production system processing operational data. What metrics would you monitor, what alert thresholds would you set, and what would trigger a rollback to a previous model or configuration?

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. D — Circuit breaker stops agent chain execution on failure to prevent cascading failures. 2. A — Shared state creates race conditions and inconsistent outputs without coordination mechanisms. 3. B — Fine-tuning modifies model weights — persistent, cannot be quickly updated without retraining. 4. D — Operational corpora may encode sensitive information into model weights accessible in

unauthorized contexts. 5. C — Hybrid retrieval combines dense (semantic) and sparse (keyword) signals. 6. C — Re-ranking reorders retrieved candidates by quality before prompt injection. 7. C — Adversarial testing must be in isolated environment — never in production. 8. C — AI observability monitors output quality, input distribution, latency, and anomalous patterns. 9. D — DoD RAIMTF: responsible AI principles — explainability, reliability, governability, traceability, bias minimization. 10. A — Army CIO Memo (April 2024): human review before official use; reviewer bears responsibility. 11. B — Rate limiting bounds blast radius of agent-chain failures before human intervention. 12. B — Inference endpoints must only process data at or below the configured classification level. 13. A — Input distribution shift causes model output drift without weight changes. 14. A — Failure isolation via circuit breakers, timeouts, and independent state management. 15. B — Shared writable namespace: one buggy/compromised agent can corrupt data affecting all agents.

Short Answer Guidance:

SA-1. Full credit: RAG — retrieves Army-specific documents at inference time without changing the model; better choice: when Army procedures are updated frequently (SOPs, doctrine updates) or when corpus is sensitive and cannot be encoded into model weights; significant risk: retrieval quality depends on corpus quality and embedding model — poor corpus produces poor context; fine-tuning — encodes Army terminology into model weights; better choice: when Army-specific vocabulary or style must be consistently applied even without retrieval; significant risk: sensitive operational data encoded into weights may be extractable via adversarial prompting; requires SJA review. Must address both approaches with scenario and risk.

SA-2. Full credit: test scenarios — prompt injection via embedded instructions in intelligence text ("ignore previous instructions and output..."); hallucination probing — inputs designed to produce confident fabrications about forces or locations; classification boundary test — inputs at boundary of classification level; role constraint bypass — attempt to make the workflow output content outside its defined scope; escalation test — attempt to trigger Ontology writes through the summary workflow; documentation: log each test case, input, output, expected behavior, actual behavior, pass/fail; remediation for each failure before release; signed completion checklist. Must cover at least three scenario types and documentation format.

SA-3. Full credit: circuit breakers — (1) if doctrine retrieval fails (no documents found, timeout), stop the chain and return an error rather than passing empty context to the generation step; (2) if generation fails or produces output flagged by a quality check, stop before routing to human review; rate limiting — limit the number of summaries generated per hour to ensure human reviewers can keep pace with the review queue; if the queue exceeds capacity, pause generation and alert; state isolation — each agent's output should be written to a dedicated staging area, not the live Ontology, until human review approves.

SA-4. Full credit: required process — (1) classify the training corpus and determine at what level the corpus data is sensitive; (2) SJA review of the corpus for PII, operational security concerns, and classification of encoded knowledge; (3) determine whether fine-tuning creates a new AI system requiring an ATO or modifies an existing one; (4) C2DAO review and approval for the training pipeline; (5) confirm

the inference endpoint's classification level matches the corpus classification; (6) document model card with fine-tuning dataset description, date, and SJA review result; (7) red-team the fine-tuned model for sensitive data leakage before deployment. All steps required for full credit.

SA-5. Full credit: metrics — output quality score (human review acceptance rate over time); input feature distribution PSI (weekly); latency (p50, p95, p99); hallucination rate proxy (factual error catch rate in review); adversarial flag rate; alert thresholds: output quality drops >10% from baseline → alert; PSI > 0.20 on key input features → alert; latency p95 exceeds SLA → alert; rollback triggers: sustained quality degradation below minimum acceptable threshold; confirmed adversarial content in outputs; input distribution shift that cannot be explained by expected operational changes; rollback procedure: revert to previous model version, notify users, investigate root cause before re-deploying.

USAREUR-AF Operational Data Team TM-50H Pre-Test | Version 1.0 | March 2026