PUBLICATION

EXAM-TM50M-POST

POST-TEST — SL 5M: ADVANCED ML 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

POST-TEST — SL 5M: ADVANCED ML ENGINEER

MAVEN SMART SYSTEM (MSS) — USAREUR-AF

Field	Detail
Course	SL 5M: Advanced ML Engineer
Form	Post-Test
Level	SL 5M (Advanced Specialist)
Audience	Senior ML engineers / data scientists; prerequisite: SL 4M + production ML experience
Time Allowed	45 minutes
Passing Score	70% (42/60)

INSTRUCTIONS

This assessment evaluates mastery of course learning objectives. A passing score of 70% is required to receive credit. Complete independently without reference to training materials.

SECTION 1 — MULTIPLE CHOICE

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

1. A SL 5M automated model promotion gate requires C2DAO approval before promoting to production. This means:

A. The pipeline automatically promotes the model if C2DAO's automated checklist passes B. C2DAO approval is only required for models that process SECRET data C. A data steward and C2DAO review must occur as a human-in-the-loop step — the promotion is not automatic regardless of gate results D. C2DAO approval replaces the standard ML acceptance threshold check

2. A new vehicle failure prediction model achieves higher AUC-ROC than the current production model. The decision to promote to production should be based on:

A. A comprehensive evaluation against all acceptance thresholds (recall ≥75%, calibration, AUC-ROC, disaggregated subgroup performance) plus C2DAO review — one metric improvement is not sufficient B. AUC-ROC improvement alone — it is the definitive production readiness metric C. A GO-level approval based on the AUC-ROC improvement D. The model's performance on the training set, which demonstrates its capability ceiling

3. In a canary release where 10% of traffic is routed to a new model version, a statistically valid comparison requires:

A. Running the canary for a minimum of 24 hours B. Running the canary until a sufficient sample size is reached to detect the minimum operationally meaningful difference at the required statistical power C. Comparing the canary model's accuracy to the full production model's training accuracy D. Running both models on the same traffic simultaneously using request duplication

4. Cross-domain federated learning between USAREUR-AF and a NATO partner requires which approvals BEFORE any technical integration begins?

A. Technical approval from both organizations' IT departments B. Army CIO approval for cross-coalition ML programs C. SJA review only — federated learning is covered under existing data sharing authorities D. C2DAO approval, both organizations' data steward sign-off, and a signed information sharing agreement covering the federated model gradient exchange — the technical claim "no data moves" does not eliminate governance requirements

5. A SHAP analysis of a failure prediction model shows that `unit_echelon` has the highest global feature importance but should be operationally irrelevant to equipment failure. This finding should trigger:

A. Removing the feature from the model and retraining B. An investigation for potential spurious correlation or proxy discrimination — the feature may be correlated with a confounding variable; review whether its inclusion introduces bias before deciding whether to remove it C. Documenting the finding in the model card without further action D. Retraining the model without `unit_echelon` and confirming AUC-ROC does not decrease

6. A disaggregated evaluation finds that the failure prediction model has 71% recall for wheeled vehicles but only 52% recall for tracked vehicles. Per USAREUR-AF standards, this means:

A. The model is acceptable — 71% exceeds the 70% threshold for the dominant vehicle type B. The model fails the acceptance criteria — both subgroup recalls must meet the ≥75% recall threshold; the tracked vehicle subgroup requires investigation and retraining C. The model is acceptable for wheeled vehicles only — deploy with a warning label for tracked vehicles D. Report the finding to the G4 and let them decide whether the discrepancy is acceptable

7. Model compression (quantization to int8) reduces the model's AUC-ROC from 0.87 to 0.83. Per SL 5M standards, the correct action is:

A. Evaluate the compressed model against ALL acceptance thresholds independently; the compressed model is a different model and prior production approval does not transfer B. Accept the degradation — 0.83 is still above 0.80, which is a common threshold C. Re-compress at float16 precision to find a compression level that maintains AUC-ROC D. Submit a waiver to the data steward documenting the 4-point AUC-ROC degradation

8. The "PSI threshold documentation" requirement in SL 5M means:

A. Recording the PSI threshold value in the pipeline configuration comments B. Setting the PSI threshold at the Foundry platform level for all models in the project C. Documenting the chosen threshold, the rationale for it (risk tolerance for this model and use case), and the escalation procedure when it is crossed — creating an auditable governance record D. Reviewing and updating PSI thresholds annually regardless of model performance

9. An adversarially robust model should be evaluated using which specific attack type for a failure prediction system?

A. Adversarial example attacks that perturb feature values within plausible operational ranges to test whether small input changes produce dramatically different failure predictions B. Black-box model extraction attacks only — white-box attacks require internal access C. SQL injection attacks against the feature computation pipeline D. Prompt injection attacks against the model serving endpoint

10. A "model registry" in a SL 5M MLOps architecture must record which of the following for each registered model version?

A. Model source code and hyperparameter search results only B. Model artifact and the date it was promoted to production C. Model artifact hash, training data version, performance metrics on validation set, calibration results, disaggregated subgroup metrics, C2DAO approval status, and deployment history D. Training compute cost and inference latency benchmarks

11. "Real-time inference" latency optimization for a vehicle failure prediction model deployed on Foundry should first address:

A. Compressing the model weights to reduce memory footprint B. Profiling the serving pipeline to identify the bottleneck (feature computation, model inference, or Ontology write) before applying any optimization C. Reducing the number of features to shorten the feature vector computation time D. Switching to a lighter model architecture without evaluating accuracy impact

12. The "no data moves" claim in a federated learning proposal is insufficient for governance approval because:

A. Data always moves in federated learning — the claim is technically incorrect B. Regulatory frameworks require that all data exchanges be explicitly authorized regardless of form C. The claim applies only to the raw data, not the model artifacts that are shared D. Model gradients can encode information about the training data and may be reconstructed via gradient inversion attacks, meaning sensitive information can be exposed through the gradient exchange

13. A neural network model for operational text classification (e.g., categorizing maintenance SITREPs) is retrained on 6 months of new data. Before deploying the updated model, you must:

A. Verify the new model achieves higher accuracy than the previous version B. Submit the model to the MSS program office for technical review C. Confirm the model's training loss converged and the validation loss did not increase D. Run the full acceptance evaluation suite including all acceptance thresholds, calibration check, disaggregated evaluation, and adversarial robustness test — and obtain C2DAO approval before promotion

14. A feature store's "online store" component serves which function that the "offline store" does not?

A. Providing low-latency feature retrieval for real-time inference requests — the offline store is optimized for batch processing, not sub-millisecond response times B. Versioning feature definitions for auditability C. Storing the complete feature history for model retraining D. Computing complex features that require access to the full historical record

15. The SL 5M "peer review by a second SL 5M qualified engineer" requirement for production ML work exists because:

A. Two engineers working independently reduces the probability of bugs by 50% B. Army policy requires two-person integrity for all production code deployments C. The complexity and operational stakes of SL 5M work (architecture decisions, adversarial robustness, bias auditing, production promotion gates) require a second expert reviewer to catch methodological errors, governance gaps, or security issues not visible to a single engineer D. Peer review replaces the C2DAO approval step for ML model promotions

---

SECTION 2 — SHORT ANSWER

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

SA-1. Design the complete MLOps pipeline for a vehicle failure prediction model from training through production deployment. Include the automated promotion gates, C2DAO approval step, canary release procedure, and the conditions that would trigger rollback.

SA-2. You have been asked to evaluate whether the USAREUR-AF vehicle failure prediction model contains bias that would result in certain units receiving systematically worse predictions. Describe your complete bias audit methodology, including the subgroups, metrics, thresholds,

and what you would do if bias is detected.

SA-3. A proposal is made to federate ML model training between USAREUR-AF and a partner nation using privacy-preserving federated learning. Walk through every governance approval required before technical integration can begin and explain why the "no data moves" claim does not satisfy the governance requirement.

SA-4. After deploying a new SITREP classification model to production, you notice the model's performance has degraded over 3 months without any code changes. Walk through your root cause investigation: what data would you examine, what metrics would you compute, and what is the most likely explanation and fix?

SA-5. Describe the SL 5M platform architecture for the USAREUR-AF ML platform, including the feature store, model registry, experiment tracking, and shared ML infrastructure components. Explain how this architecture prevents the three most common failure modes in production ML systems.

SCORING SUMMARY

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

Passing: 42/60 (70%) — Post-test only. Pre-test is diagnostic.

ANSWER KEY — INSTRUCTOR USE ONLY

Do not distribute to students.

Multiple Choice: 1. C — C2DAO approval is a human-in-the-loop step, not an automated gate bypass. 2. A — All acceptance thresholds (recall, calibration, AUC, disaggregated) plus C2DAO review required. 3. B — Statistically valid canary requires sufficient sample size for the minimum meaningful difference at required power. 4. D — C2DAO + data stewards + ISA required before any technical integration; "no data moves" is insufficient. 5. B — Unexpected feature importance requires investigation for spurious correlation or proxy discrimination. 6. B — 52% recall for tracked vehicles fails the ≥75% threshold; the model cannot be deployed as-is. 7. A — Compressed model must be independently evaluated against all acceptance thresholds. 8. C — PSI threshold documentation = threshold value + rationale + escalation procedure, creating governance record. 9. A — Adversarial examples with plausible feature perturbations test failure prediction stability. 10. C — Model registry must record artifact hash, training data version, metrics, calibration, disaggregated metrics, C2DAO status, deployment history. 11. B — Profile the pipeline to identify the bottleneck before applying optimizations. 12. D — Gradients can encode training data information; gradient inversion attacks can reconstruct sensitive data. 13. D — Full acceptance suite + calibration + disaggregated + adversarial test + C2DAO approval required. 14. A — Online store provides low-latency real-time feature retrieval for inference requests. 15. C — SL 5M complexity and operational stakes require second expert reviewer for errors not visible to one engineer.

Short Answer Guidance:

SA-1. Full credit: pipeline stages — (1) training with versioned dataset and pinned dependencies; (2) automated evaluation gate: check recall ≥75%, calibration, AUC ≥threshold, disaggregated subgroup metrics — all must pass; (3) C2DAO review (human-in-the-loop, not automated); (4) staging deployment and integration testing; (5) canary release (10% traffic) with statistical monitoring for minimum detectable

degradation; (6) full production cutover if canary passes; rollback conditions: recall drops below 75% in production, PSI exceeds threshold on key features, or canary detects statistically significant performance degradation vs. current model. Must include all six stages and specific rollback conditions.

SA-2. Full credit: subgroups — vehicle type (wheeled, tracked, aviation), unit echelon (company, battalion, brigade), vehicle age cohort (0–3 yrs, 3–7 yrs, 7+ yrs), geographic region (AOR); metrics — recall per subgroup, false negative rate per subgroup, calibration per subgroup; thresholds — any subgroup recall < 75% = model failure; any subgroup FNR > 2x overall FNR = potential bias; if bias detected: (1) investigate root cause — is the subgroup underrepresented in training data? (2) oversample the underperforming subgroup; (3) add subgroup-specific features; (4) retrain and re-evaluate; document findings in model card regardless of outcome.

SA-3. Full credit: required approvals in order — (1) USAREUR-AF C2DAO review and approval for the federated learning architecture; (2) USAREUR-AF data steward sign-off on the data governance implications; (3) partner nation equivalent governance approval; (4) signed information sharing agreement (ISA) covering gradient exchange; (5) SJA review of the ISA and classification handling; (6) legal review of the partner nation's privacy and data protection requirements; why "no data moves" is insufficient: model gradients can encode information about training examples — gradient inversion attacks can reconstruct individual training records from gradient updates; the sensitive information moves in the gradients even if raw data does not.

SA-4. Full credit: root cause investigation — (1) check input feature PSI for all features: look for distribution shift between training period and current production inputs; (2) compute prediction score distribution PSI: has the model's output distribution shifted?; (3) compare vocabulary and content of current SITREPs to training set SITREPs — new terminology, operational changes, or reporting format changes?; (4) check for ground truth label drift if labels are available; most likely explanation: input distribution shift — SITREP language and content has evolved since training; fix: collect recent labeled examples, retrain on updated corpus, run full acceptance suite, obtain C2DAO approval, redeploy via canary.

SA-5. Full credit: components — (1) Feature store: offline store (Foundry dataset with versioned feature definitions, batch-computed, refreshed on schedule); online store (low-latency key-value store for real-time inference); shared feature library imported by both training and serving; (2) Model registry: versioned model artifacts with metadata (training data version, metrics, calibration, C2DAO status); (3) Experiment tracking: logs all training runs (hyperparameters, metrics, artifacts) for reproducibility and comparison; (4) Shared infrastructure: GPU workspace templates, standard pipeline templates, shared evaluation functions; three common failure modes prevented: (1) training-serving skew → prevented by shared feature store and library; (2) silent model degradation → prevented by drift monitoring pipeline and PSI alerting; (3) ungoverned model promotion → prevented by automated evaluation gates and C2DAO approval requirement.

USAREUR-AF Operational Data Team TM-50M Post-Test | Version 1.0 | March 2026