PUBLICATION

EXAM-TM50M-PRE

PRE-TEST — SL 5M: ADVANCED ML ENGINEER

Maven Smart System (MSS) — USAREUR-AF

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

DRAFT — NOT FOR OFFICIAL USE. FOR TRAINING PLANNING PURPOSES ONLY.

26 MARCH 2026

DRAFT — UNOFFICIAL — NOT FOR OPERATIONAL USE

PRE-TEST — SL 5M: ADVANCED ML ENGINEER

MAVEN SMART SYSTEM (MSS) — USAREUR-AF

Field	Detail
Course	SL 5M: Advanced ML Engineer
Form	Pre-Test
Level	SL 5M (Advanced Specialist)
Audience	Senior ML engineers / data scientists; prerequisite: SL 4M + production ML 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 production MLOps pipeline, "automated model promotion gates" control:

A. The conditions that must be met before a trained model artifact is promoted from development to staging to production B. The schedule at which models are retrained on new data C. The access permissions for who can deploy models D. The compute allocation for production model inference

2. A "canary release" for a new ML model version involves:

A. Deploying the new model to all users simultaneously and monitoring for problems B. Routing a small percentage of production traffic to the new model while the majority runs on the current model, enabling comparison before full rollout C. Testing the new model on a canary dataset held out from the original training set D. Running the new model in shadow mode without returning its predictions to users

3. "A/B testing" for ML models in production measures:

A. Whether model A was trained on dataset A and model B on dataset B B. The latency difference between two model architectures on the same hardware C. Whether the model's accuracy is significantly above random chance using a t-test D. The difference in a defined metric (accuracy, conversion rate, operational outcome) between two model versions when exposed to equivalent real-world traffic

4. "Federated learning" is a training approach where:

A. Multiple organizations jointly fund model training on a shared centralized dataset B. The model is trained in a federated government cloud environment with shared compute C. Models are trained locally on distributed data sources without transferring the raw data to a central location — only model updates are shared D. Multiple model architectures are trained in parallel on the same dataset

5. Cross-domain federated learning coordination (e.g., between USAREUR-AF and a partner nation) requires:

A. C2DAO approval, data steward sign-off, and a signed information sharing agreement — the claim "no data moves" is insufficient because model gradients may encode sensitive information B. Technical integration only — federated learning inherently provides privacy by design C. Approval from the partner nation's equivalent of the Army CIO D. SJA review only for SECRET or above data

6. "SHAP (SHapley Additive exPlanations) values" provide:

A. A measure of global model accuracy across all predictions B. A statistical test for whether the model is significantly better than a baseline C. Feature-level attributions showing how much each feature contributed to an individual prediction D. The marginal effect of each feature after controlling for correlations

7. "Disaggregated evaluation" in ML fairness assessment means:

A. Evaluating model performance separately on different demographic or operational subgroups to detect performance disparities B. Separating the model's training data into subsets for multi-fold cross-validation C. Evaluating each feature independently to identify which features drive bias D. Running the model on disaggregated data from multiple time periods

8. "Data poisoning" in an adversarial ML context is a training-time attack where:

A. Malicious training examples are injected to manipulate the model's learned behavior in a targeted or indiscriminate way B. Corrupt data is introduced into the production inference pipeline C. The model's gradient updates are intercepted and modified during distributed training D. The training dataset is intentionally biased by removing representative samples

9. "Model extraction" attacks attempt to:

A. Reconstruct or approximate a deployed model's behavior by querying it systematically, enabling the attacker to build a functionally equivalent copy B. Steal the model's training data by querying the model extensively C. Extract the model's weights from a compromised serving endpoint D. Modify the model's outputs by manipulating the inference endpoint

10. "Post-training quantization" in model compression involves:

A. Retraining the model with a lower learning rate after initial convergence B. Reducing the numerical precision of model weights (e.g., float32 to int8) after training to reduce memory footprint and inference latency C. Removing low-importance features from the input before inference D. Compressing the model using knowledge distillation from a larger teacher model

11. A "feature store" in an ML platform serves which primary purpose?

A. A version-controlled repository for model artifacts B. A catalog of approved features for each ML use case C. A database for storing raw training data separate from production data D. A centralized system for computing, storing, and serving features consistently across training and inference environments, preventing training-serving skew

12. "Training-serving skew" in an ML deployment occurs when:

A. The model architecture used in training differs from the serving architecture B. The model's output schema in training differs from the output schema at inference C. The training dataset and serving dataset come from different time periods D. The feature computation logic differs between the training pipeline and the online inference pipeline, causing the served model to receive different feature distributions than it was trained on

13. "Bias auditing" of a vehicle failure prediction model would specifically examine:

A. Whether the training dataset is large enough to avoid overfitting B. Whether the model's false negative rate is significantly higher for specific unit types, vehicle ages, or geographic regions — indicating that certain subgroups are systematically underserved C. Whether the model's feature importances align with domain expert expectations D. Whether the model is more accurate than a rules-based baseline

14. The "PSI threshold" for triggering automated model retraining review should be documented because:

A. The Foundry platform requires PSI thresholds to be registered in the pipeline configuration B. PSI thresholds are fixed at 0.20 by Army ML policy C. Different models and use cases warrant different sensitivity to input distribution shift — the chosen threshold reflects a deliberate risk tolerance decision that should be transparent and auditable D. Documentation prevents the threshold from being inadvertently changed during pipeline updates

15. A "graph neural network" (GNN) is appropriate for modeling operational C2 networks because:

A. C2 networks have a fixed structure that requires specialized graph embeddings B. GNNs are more interpretable than standard neural networks for military use cases C. GNNs naturally represent and learn from relational structure — nodes (units, commanders) and edges (command relationships,

communication links) — enabling predictions that account for network topology D. C2 network data cannot be flattened into a tabular format

SECTION 2 — SHORT ANSWER

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

SA-1. Describe the difference between a canary release and a shadow deployment for an ML model. For each, describe a specific operational scenario in an Army data context where you would use it and why.

SA-2. Explain what "training-serving skew" is and describe two concrete ways it could arise in an MSS ML deployment. How would you detect it, and what is the fix?

SA-3. You are asked to audit a vehicle failure prediction model for bias. Describe your methodology: what subgroups you would evaluate, what metrics you would compute, and what your decision criteria would be for determining whether the model has an unacceptable bias.

SA-4. Describe the key security risks of deploying an ML model in an adversarial operational environment. For each risk, describe one defensive control.

SA-5. Explain what a "feature store" is, how it prevents training-serving skew, and describe how you would design a feature store for the USAREUR-AF predictive maintenance ML platform.

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. A — Automated promotion gates define conditions for model progression through environments. 2. B — Canary release routes small traffic percentage to new model for comparison. 3. D — A/B testing measures defined metric differences between two model versions on real traffic. 4. C — Federated learning trains locally on distributed data; only updates (not raw data) are shared. 5. A — Cross-domain federated learning requires C2DAO + data steward + ISA; "no data moves" is insufficient. 6. C — SHAP provides per-feature attributions for individual predictions. 7. A — Disaggregated evaluation assesses performance separately on different demographic or operational subgroups to detect performance disparities. 8. A — Data poisoning = adversarial training example injection to manipulate learned behavior. 9. A — Model extraction = systematic querying to reconstruct model behavior. 10. B — Post-training quantization reduces weight precision after training to reduce memory/latency. 11. D — Feature store provides consistent feature computation across training and inference. 12. D — Training-serving skew = different feature computation logic in training vs. serving pipelines. 13. B — Bias audit

checks for false negative rate disparities across subgroups. 14. C — PSI threshold documents a deliberate, auditable risk tolerance decision. 15. C — GNNs learn from relational structure of nodes and edges representing C2 topology.

Short Answer Guidance:

SA-1. Full credit: canary release — routes a fraction (5–10%) of production traffic to new model while most traffic uses current model; use when: deploying a new failure prediction model version where you need to compare real prediction outcomes before full cutover (parallel production comparison with live outcomes); shadow deployment — new model runs on production inputs but outputs are not served to users — only logged for comparison; use when: validating a major model architecture change where serving incorrect predictions would have operational consequences, and you need validation before any real-world exposure. Both scenarios must be Army-context specific.

SA-2. Full credit: definition — feature computation logic differs between training and serving, causing the model to receive different distributions at inference than during training; two examples: (1) days_since_service computed using CURRENT_DATE in training snapshots but using the record's snapshot_date at inference — different baseline; (2) a categorical feature encoded as one-hot in training but label-encoded in the serving pipeline; detection: compare feature distribution statistics (mean, variance, PSI) between training data and production inference inputs; fix: use a shared feature store or shared feature computation library that both training and serving pipelines call.

SA-3. Full credit: subgroups — vehicle type (wheeled vs. tracked), unit echelon (brigade vs. battalion), vehicle age cohort, geographic region; metrics — false negative rate (missed failures) per subgroup, and comparison to overall false negative rate; decision criteria: if any subgroup's false negative rate exceeds overall FNR by >15 percentage points (or a defined operational threshold), model has unacceptable bias; additionally check recall per subgroup against the ≥75% USAREUR-AF minimum threshold — any subgroup falling below is a failure. Partial credit (3 pts) for subgroups and metrics without decision criteria.

SA-4. Full credit: any four risks with controls — (1) data poisoning: defensive control = training data provenance tracking and outlier detection in training inputs; (2) model extraction: control = query rate limiting and output perturbation; (3) adversarial examples at inference: control = input validation and out-of-distribution detection; (4) training pipeline compromise: control = code review, pipeline integrity checks, and audit logs; (5) model artifact tampering: control = model registry with hash verification. Each risk must have a specific control.

SA-5. Full credit: feature store = centralized system for computing, storing, and serving features consistently so training and inference pipelines use identical computation logic; prevents training-serving skew by providing a single source of truth for feature definitions; design for USAREUR-AF predictive maintenance — offline store: batch computation of historical features (days since service, failure count, mileage) stored in Foundry dataset, refreshed on pipeline schedule; online store: low-latency feature retrieval for real-time inference (if used); feature definitions versioned: any change to a feature definition

creates a new version and triggers retraining evaluation; shared library: a single Python library that both training and serving pipelines import for feature computation. Must include offline/online distinction and versioning for full credit.

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