Sprint 4 Completion Report: LLM Egress Guard

Project: LLM Egress Guard - Data Loss Prevention for LLM Outputs
Sprint: Sprint 4 (Dec 1 - Dec 15, 2025)
Status: [COMPLETE]
Prepared by: Baran Akin
Date: Dec 15, 2025

Executive Summary

Sprint 4 delivers a functional ML Pre-Classifier (TF-IDF + Logistic Regression), integrated behind a feature flag, with shadow/A-B instrumentation and model integrity checks. Context-aware parsing and FP reduction from Sprint 3 remain stable. Grafana dashboards are deferred to Sprint 5.

Objectives & Status

Objective	Description	Status
ML Pre-Classifier v1	Train TF-IDF + Logistic Regression on synthetic data	✅ Complete
ML Integration (feature flag)	Load model via env; heuristic fallback	✅ Complete
Shadow/A-B Instrumentation	Log ML vs heuristic disagreements	✅ Complete
Model Manifest & Checksum	Manifest + verification script	✅ Complete
Context-aware parsing & FP reduction	Carried over, stable	✅ Complete
Grafana dashboards	Prometheus panels for ML & context metrics	⏭ Deferred → Sprint 5

Implementation Highlights

- **Model:** `models/preclf_v1.joblib`
  - Eval: Accuracy 0.8857; F1 macro 0.8604
  - Per label: educational f1=0.923, command f1=0.889, text f1=0.769
- **Manifest:** `models/preclf_v1.manifest.json` (sha256, size recorded)
- **Loader:** `app/ml/preclassifier.py` loads model; falls back to heuristic on failure
- **Pipeline:** `FEATURE_ML_PRECLF` + `PRECLF_MODEL_PATH` control model use; `SHADOW_MODE` logs ML vs heuristic diffs without changing decisions
- **Metrics:**
  - `egress_guard_ml_preclf_load_total{status}` (success/fail)
  - `egress_guard_ml_preclf_shadow_total{ml_pred,heuristic,final}` (disagreements)
  - Existing context/blocked/rule metrics remain
- **Validation script:** `scripts/check_preclf_model.py` verifies checksum vs manifest

Configuration

```
# Enable ML pre-classifier
export FEATURE_ML_PRECLF=true
export PRECLF_MODEL_PATH=models/preclf_v1.joblib

# Optional: shadow mode (logs ML vs heuristic differences; no
        decision change)
export SHADOW_MODE=true
```

Testing & Metrics

- Synthetic dataset: 175 samples → Train 140 / Eval 35
- Training command:

```
python scripts/train_preclassifier.py \
  --train data/ml_training/preclf_train.jsonl \
  --eval data/ml_training/preclf_eval.jsonl \
  --output models/preclf_v1.joblib
```

- Model verification:

```
python scripts/check_preclf_model.py \
  --model models/preclf_v1.joblib \
  --manifest models/preclf_v1.manifest.json
```

- Smoke tests (API):
  - Educational tutorial with `curl|bash` → blocked=false, explain_only=true
  - Malicious `curl|bash` instruction → blocked=true

Deferred to Sprint 5

Item	Reason
Grafana dashboards for ML/context metrics	Requires dashboard setup (external)

Summary

Sprint 4 completes ML pre-classifier training, integration, observability (load/shadow metrics), and model integrity checks. The system is ready to run ML in production behind a feature flag, with shadow mode available for safe comparison to heuristic logic. Remaining task (dashboards) is scheduled for Sprint 5.

Sprint 4 Status: □ COMPLETE
Next (Sprint 5): Dashboards and any tuning based on shadow-mode findings.