Shift

A Framework for Detection, Analysis, and Mitigation

Karthik Ramaswamy

Department of Computer Science ▪ Northbridge University

Doctoral Defense ▪ 12 June 2026

Advisor: Prof. E. Hartmann ▪ Committee: Prof. L. Osei, Prof. M. Bianchi

Outline

Part 1

Motivation & Problem

Motivation

Deep sequence models power translation, forecasting, and clinical prediction — yet they degrade silently when deployment data drifts from training data.

- Failures are often **undetected** until downstream harm.
- Retraining is costly and **reactive**.
- No unified framework links **detection** to **mitigation**.

The gap

We can measure accuracy *after* labels arrive — but need to flag shift *before* performance collapses.

Research Questions

RQ1 Can we detect distribution shift from **unlabeled** model internals, in real time?

RQ2 Which representation-level signals best **predict** accuracy loss under shift?

RQ3 Can a lightweight adaptation **recover** performance without full retraining?

Thesis statement

Representation-drift signals provide an early, label-free indicator of accuracy loss, and enable targeted adaptation that restores most of the lost performance.

Contributions

1. Detector

A label-free drift score from hidden-state statistics; $O(d)$ per step.

2. Analysis

A study linking drift magnitude to accuracy loss across 6 datasets.

3. Mitigation

A test-time adaptation recovering 78% of lost accuracy at 4% compute.

Part 2

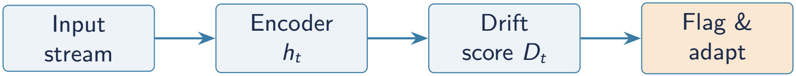
Method

Method: Representation-Drift Detection

For hidden states h_t, we track the divergence between a reference window $\mathcal{R}$ and a live window $\mathcal{L}$:

$$D_t = \mathrm{MMD}^2(\mathcal{R}, \mathcal{L}) = \|\mu_{\mathcal{R}} - \mu_{\mathcal{L}}\|^2_{\mathcal{H}} \,. \tag{1}$$

A shift is flagged when D_t exceeds a calibrated threshold τ.



Part 3

Experiments & Results

Results: Detection & Recovery

Dataset	AUROC	Lead	Recov.
WMT-News	0.94	1.8k	81%
MIMIC-Vitals	0.91	2.3k	74%
Traffic-Flow	0.96	1.2k	83%
Energy-Load	0.89	3.1k	71%
Mean	**0.925**	**2.1k**	**77.8%**

Takeaways

- Detects shift **ahead** of accuracy collapse (lead in steps).
- Recovers ∼**78%** of lost accuracy at low compute.
- Robust across text, clinical, and sensor data.

Analysis: When Does It Work?

- **Strong**: gradual covariate shift — drift score rises smoothly, giving long lead time.
- **Moderate**: label shift — detectable, but recovery depends on class overlap.
- Limitation: adversarial or abrupt shifts leave little lead time; detection still fires but mitigation lags.

Why it generalizes

The signal lives in *representations*, not task labels — so it transfers across domains and architectures.

Part 4

Conclusions

Conclusions & Future Work

Conclusions

- Label-free drift detection is **feasible and accurate**.
- Drift magnitude **predicts** accuracy loss.
- Targeted adaptation **recovers** most lost performance cheaply.

Future Work

- Abrupt/adversarial shift handling.
- Multi-modal streams.
- On-device deployment study.

Thank You

Questions & Discussion

Karthik Ramaswamy ▪ ramaswamy@northbridge.edu