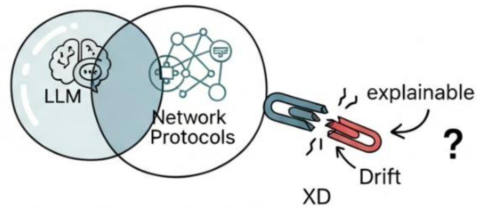


XD: eXplainable Drift for LLM’s Robustness in Communication Network Protocols Modelling

Abhishek Djeachandrane*, Jorge Lopez*, Charalampos Chatzinakis*

* Airbus, Boulogne-Billancourt, France

AIRBUS

Lyon, France, 17-22 August 2026
International Conference on Pattern Recognition

ICPR 2026 International Conference on Pattern Recognition LYON

Outline

- Introduction: Context-Problem-Solution
- Proposed Methods
- Experiment Results
- Conclusion and Perspectives

Introduction

Context - Critical Communication Network

- **Conventional Network**
  - Environment ➔ Network Infrastructure / Topology (Data Plane, links, routers)
  - Intelligence Layer ➔ Control Plane / Routing Protocols (e.g., OSPF, SDN Controller)
  - Algorithms ➔ Optimization Techniques (e.g., Dijkstra's algorithm, Linear Programming)
  - Decision Variables ➔ Routing Parameters & Rules (e.g., Link Weights, Next-Hop Tables)

Traditional Network =/ Critical Network

Critical by **strict time constraints** & **bad decision consequences** …

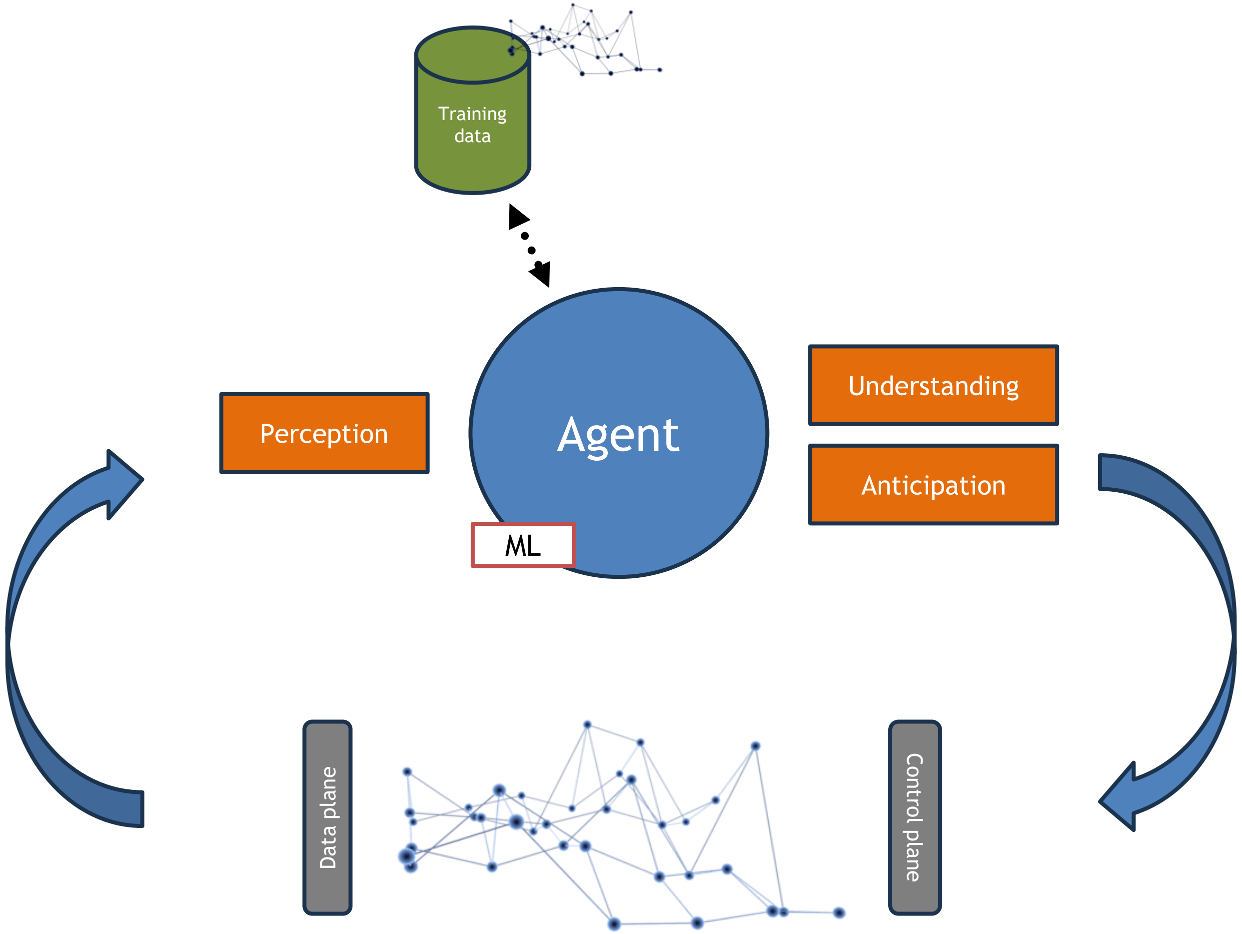
- **For critical application,**
  - Look for better "Relevant" decision variables ;
  - Anticipation is key rather than Reaction in near real-time application.

- **Technological trends x Operational realities:**
  - Literature review : LLM is widely used in Telecommunication;
  - Operational realities : For network usage monitoring, application protocols are raising greater concerns.

AIRBUS

Introduction

High-Level Representation of our System and Task



Task : Predict the network protocols at t+1 with reliability and transparency

Introduction

Problem statement

Trustworthy Informed Prediction for Critical Network Control

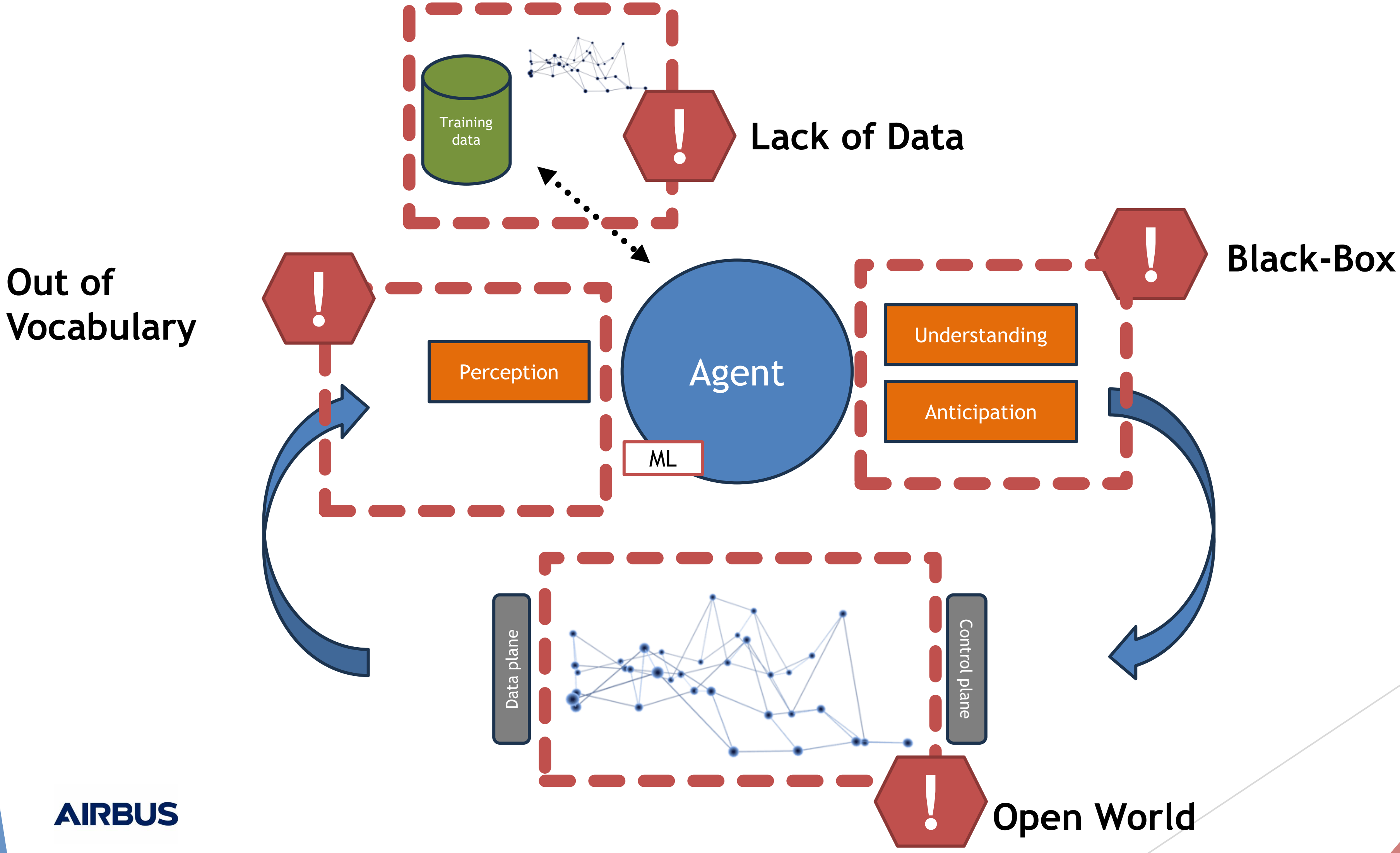
How does the system operate and be accurate in an uncontrolled critical environment ?

Perception of the environment;	Understanding of the environment;	Anticipation of the environment.

Knowledge Validation

Introduction

High level representation of our System & Locks



AIRBUS

Introduction

From bottlenecks to solutions through established concepts

Concepts

1. NLP : Representation learning
2. Frugal AI : Transfer Learning

3. Trustworthy AI
   - Robust AI
   - eXplainable AI

Locks

1. Out of Vocabulary
2. Lack of Data

3. Open World
4. Black-box

Solutions

LLM

1. Tokenizer
2. Pre-trained Model

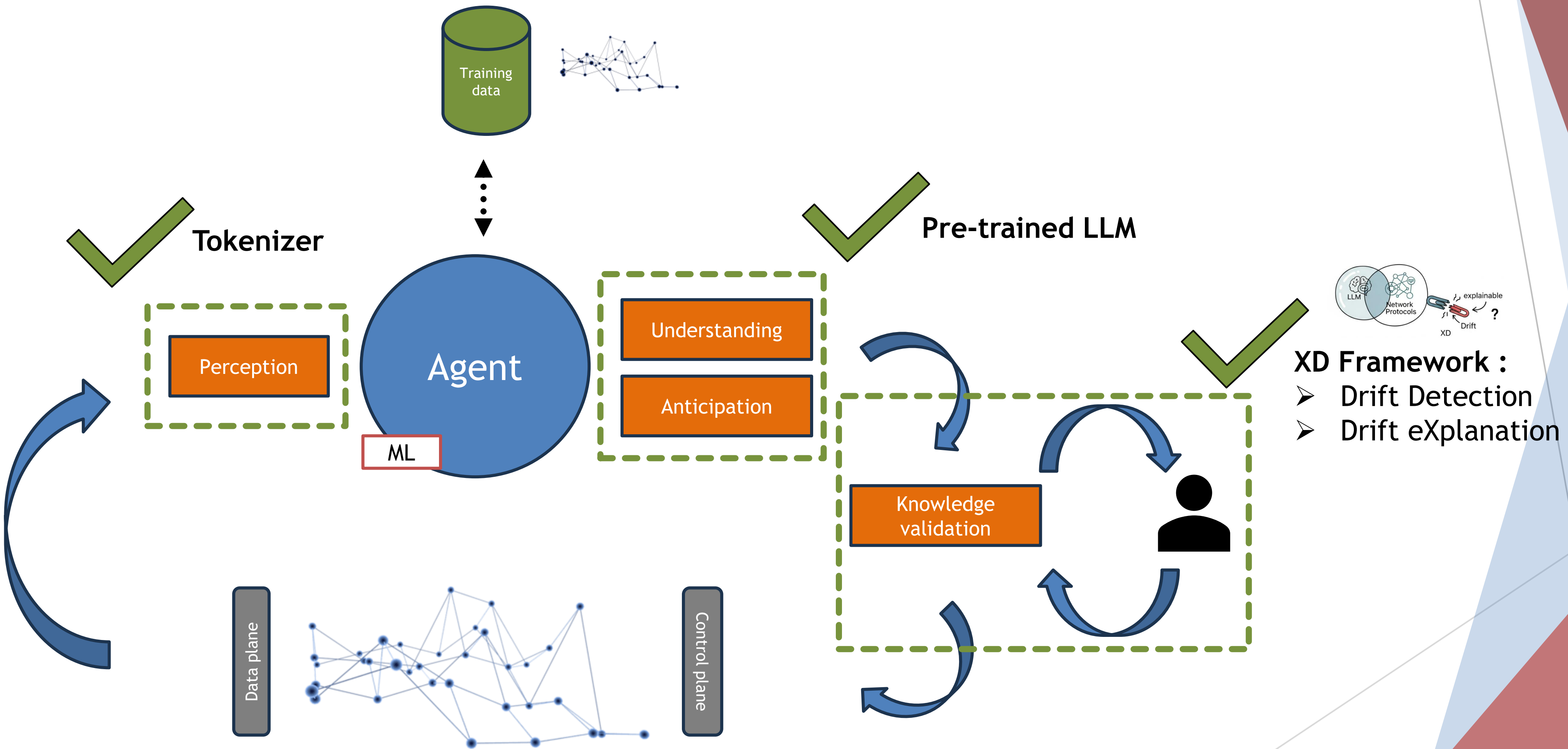
XD Framework

3. Concept and Data drift Detection
4. User aligned Drift Explanation

AIRBUS

Introduction

High level representation & solutions



AIRBUS

Proposed Methods
The Master overview

Data source and processing	Generative AI for Modelling and Prediction
Multi Level Behavioral Monitoring and Drift Detection	Human Interpretable Concept and Data Drift eXplanation

AIRBUS

Proposed Methods

The Master overview

Data source and processing	Generative AI for Modelling and Prediction
Multi Level Behavioral Monitoring and Drift Detection	Human Interpretable Concept and Data Drift eXplanation

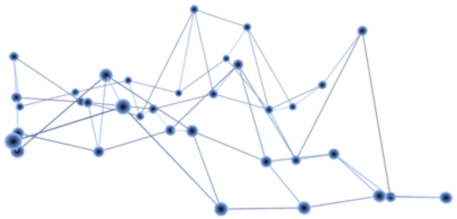
Data Source and Processing

Real World network Traffic Dataset and Simulated Data

- MAWI
  - This dataset consists of packet traces from a high-speed trans-Pacific backbone link (202304121400 trace daily transit link traffic from WIDE to its upstream ISP)
- CAIDA
  - This dataset provides large-scale traces from the Equinix-NYC commercial monitor (a 15-minute capture from January 17, 2019 (13:00:00 to 13:15:00 UTC) on high-volume commercial streams.
- UNICAUCA
  - This dataset (Version 2) focuses on enterprise-level flow-based traffic, offering a distinct protocol distribution to test the model's adaptability across network domains.
- Simulated Data
  - These sequences are produced via a Markov chain where transitions between protocols (e.g., HTTP, FTP, SSH) are governed by a configurable transition tensor. By modifying the stochastic properties of this tensor, we precisely inject shifts in protocol distribution or transition logic.

Data Source and Processing

Time Series Data Processing for Top-K Protocol Analytics



Traffic extraction → Flow aggregation → Sequence engineering

Traffic extraction

- 21 headers fields (timestamps, IP addresses Transport layer ports …) with Tshark from raw PCAP files

Flow aggregation

- Flow reconstruction identified by 5-tuple flow ID
- Temporal sampling (ranging from 5 to 20 seconds)
- enriched by application layer protocol. (HTTP, SSH …)

Sequence engineering

- Top k most frequent protocol (6,9,12)
- sequence lengths (1 to 4).

AIRBUS

Proposed Methods

The Master overview

Data source and processing

Generative AI for Modelling and Prediction

Multi Level Behavioral Monitoring and Drift Detection

Human Interpretable Concept and Data Drift eXplanation

Generative AI for Modelling and Prediction

LLM for Protocol Sequence prediction & Evaluation Metrics for Operational reliability

Category	Model	Size
Pioneer Architecture	GPT2	125M
SLM	Phi2	1.2B
	LFM2	2.7B
LLM	Mistral	7B
	Falcon	7B
	Llama-3	8B

- Protocol Level Performance Metric

$$\mathrm{Prec}_{\mathrm{protocol}} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{TP_i}{TP_i + FP_i}\right)$$

- Sequence-Level Performance Metric

$$\mathrm{Acc}_{\mathrm{sequence}} = \frac{TP_{total}}{TP_{total} + FP_{total}}$$

- Relative Performance Metrics

$$\Delta_{\mathrm{metric}} = \frac{\mathrm{Metric}_{\mathrm{model}} - \mathrm{Metric}_{\mathrm{base}}}{\mathrm{Metric}_{\mathrm{base}}} = \frac{\mathrm{Metric}_{\mathrm{model}}}{\mathrm{Metric}_{\mathrm{base}}} - 1$$

Proposed Methods

The Master overview

Data source and processing	Generative AI for Modelling and Prediction
Multi Level Behavioral Monitoring and Drift Detection	Human Interpretable Concept and Data Drift eXplanation

Multi Level Behavioral Monitoring and Drift Detection

Model Pattern Representation, Bias and Discrepancy Metrics, and Drift Detection

- Model **Pattern Representation** : $T_{uv} = \mathbb{P}(p_{t+1} = v \mid p_t = u) = \frac{\text{count}(u \to v)}{\sum_{w \in \mathcal{S}} \text{count}(u \to w)}$

- Model **Bias and Discrepancy Metrics**

  - E-wise Cross Entropy : $CE(Q,P)_{ij} = -Q_{ij} \ln(P_{ij} + \epsilon)$
  - E-wise Symmetric KL Divergence : $D_{SKL}(P,Q)_{ij} = \frac{1}{2}\left(P_{ij} \ln \frac{P_{ij}}{Q_{ij}} + Q_{ij} \ln \frac{Q_{ij}}{P_{ij}}\right)$
  - E-wise Hellinger Distance : $H(P,Q)_{ij} = \frac{1}{\sqrt{2}} \sqrt{\left(\sqrt{P_{ij}} - \sqrt{Q_{ij}}\right)^2}$

- Model **Drift Detection** : $D_t = \text{stat}\left(d(B_{gold}, B_t)\right)$ & $\bar{D}_t = \frac{1}{k} \sum_{i=0}^{k-1} D_{t-i}$

Multi Level Behavioral Monitoring and Drift Detection

Threshold Calibration Modes and Anomaly Detection Evaluation Metrics

- Threshold Calibration Modes

  - **Manual (man):** $\tau = \lambda$, where λ is a user-defined constant.
  - **Semi-Supervised (semi):** $\tau = \mu_D + (\alpha \cdot \sigma_D)$, where α is a sensitivity coefficient.
  - **Unsupervised (full):** $\tau = \mu_D + (\kappa \cdot \sigma_D)$, where κ is automatically adjusted using a buffer β:

$$\kappa = \frac{\max(D_{internal}) - \mu_D}{\sigma_D} + \beta \tag{12}$$

A drift event is signaled at time t if the smoothed drift score exceeds the calibration threshold τ: $\bar{D}_t > \tau$.

- Anomaly Detection Evaluation Metrics
  - Accuracy & Threshold-independent Area under the Curve

AIRBUS

Proposed Methods

The Master overview

Data source and processing	Generative AI for Modelling and Prediction
Multi Level Behavioral Monitoring and Drift Detection	Human Interpretable Concept and Data Drift eXplanation

Human Interpretable Concept and Data Drift eXplanation

Concept Drift Analysis, Data Drift Analysis ,and Filtering and Complexity Management

- Concept Drift Analysis (Nodes)

  – **Appearing Concepts ($\mathcal{C}_{appear}$):** Protocols that are newly observed in the current window.

  $$\mathcal{C}_{appear} = \{s \mid s \in \mathcal{S}_{curr} \land s \notin \mathcal{S}_{prev}\} \tag{13}$$

  – **Disappearing Concepts ($\mathcal{C}_{disappear}$):** Protocols that were present in the previous state but are no longer active.

  $$\mathcal{C}_{disappear} = \{s \mid s \in \mathcal{S}_{prev} \land s \notin \mathcal{S}_{curr}\} \tag{14}$$

  – **Stable Concepts ($\mathcal{C}_{stable}$):** Protocols that persist across both windows.

  $$\mathcal{C}_{stable} = \mathcal{S}_{curr} \cap \mathcal{S}_{prev} \tag{15}$$

- Data Drift Analysis (Edge)

  $$\text{Edge\_Status}(u, v) = \begin{cases} \text{Data Drift (Red)} & \text{if } B_{t,uv} > \tau_{ref} \\ \text{Stable (Gray)} & \text{otherwise} \end{cases} \tag{16}$$

- Filtering and Complexity Management
  - Node Based Filtering

    $$\Psi_{node}(u) = \sum_{v \in \mathcal{S}} |T_{uv}| + \sum_{v \in \mathcal{S}} |T_{vu}|$$

  - Link based Filtering

    $$\Psi_{link}(u, v) = |T_{uv}|$$

AIRBUS

Experiment Results

Experimental methodology step-by-step

01

Comparison Study of LLMs for Protocol Sequence Prediction

02

Offline Evaluation for Retrospective Drift Analysis

03

Online Evaluation for Real-time Drift Analysis

Experiment Results

Experimental methodology step-by-step

01	02	03
Comparison Study of LLMs for Protocol Sequence Prediction	Offline Evaluation for Retrospective Drift Analysis	Online Evaluation for Real-time Drift Analysis

Experiment Results

Comparison Study of LLMs for Protocol Sequence Prediction

Table 1. Comprehensive Experimental Results across MAWI, CAIDA, and Unicauca Datasets

Dataset	Category	Model	Task-specific (Raw)		Real-World (Relative Δ)	
			Prot-lev	Seq-lev	Prot RP (Δ)	Seq RP (Δ)
MAWI	Pioneer LM	GPT2	0.8488	0.3704	8.70%	25.00%
	SLM	Phi2	0.8272	0.4074	5.93%	37.50%
		LFM2-1.2b	0.8364	0.3704	7.11%	25.00%
	LLM	Mistral-7B	0.8272	0.3333	5.93%	12.50%
		Falcon-7b	0.8117	0.1852	3.95%	−37.50%
		Llama3-8b	0.8241	0.2963	5.53%	0.00%
CAIDA	Pioneer LM	GPT2	0.9080	0.4828	3.61%	16.67%
	SLM	Phi2	0.9167	0.5172	4.59%	25.00%
		LFM2-1.2b	0.9167	0.5172	4.59%	25.00%
	LLM	Mistral-7B	0.9023	0.4483	2.95%	8.33%
		Falcon-7b	0.9167	0.5172	4.59%	25.00%
		Llama3-8b	0.9023	0.4828	2.95%	16.67%
Unicauca	Pioneer LM	GPT2	0.4223	0.0000	11.68%	Nan
	SLM	Phi2	0.4211	0.0000	11.36%	Nan
		LFM2-1.2b	0.4186	0.0001	10.72%	Nan
	LLM	Mistral-7B	0.4178	0.0053	10.51%	Nan
		Falcon-7b	0.4113	0.0000	8.79%	Nan
		Llama3-8b	0.4101	0.0053	8.47%	Nan

Table 2. Summary of Average Real-World Performance Across Datasets

Category	Models	Av prot RP	Av seq RP w/o Unicauca
Pioneer LM	GPT2	8.00%	20.84%
SLM	Phi2	7.29%	31.25%
	LFM2-1.2b	7.47%	25.00%
LLM	Mistral-7B	6.46%	10.42%
	Falcon-7b	5.78%	−6.25%
	Llama3-8b	5.65%	8.34%

$$\Delta_{\text{metric}} = \frac{\text{Metric}_{\text{model}}}{\underbrace{\text{Metric}_{\text{base}}}_{\simeq 0}} - 1$$

AIRBUS

Experiment Results

Comparison Study of LLMs for Protocol Sequence Prediction

Table 4. Summary of Real-World Performance by Sequence Length (*Seq*)

Seq	MAWI		CAIDA		Unicauca		Averages	
	Prot RP	Seq RP	Prot RP	Seq RP	Prot RP	Seq RP	Av Prot	No Uni
1	5.14%	25.00%	−1.31%	−25.00%	8.72%	Nan	4.18%	0.00%
2	5.93%	37.50%	4.59%	25.00%	11.36%	Nan	7.29%	31.25%
3	2.77%	25.00%	−2.62%	−25.00%	17.41%	Nan	5.85%	0.00%
4	1.58%	−25.00%	2.62%	25.00%	23.60%	Nan	9.27%	0.00%

Table 3. Summary of Real-World Performance by Interval (i)

Interval	MAWI		CAIDA		Unicauca		Averages	
	Prot RP	Seq RP	Prot RP	Seq RP	Prot RP	Seq RP	Av Prot	Av Seq
i5	5.93%	37.50%	4.59%	25.00%	11.36%	Nan	7.29%	20.83%
i10	13.89%	0.00%	1.90%	60.00%	6.31%	0.00%	7.37%	20.00%
i15	−10.42%	25.00%	9.62%	28.57%	9.33%	0.00%	2.84%	17.86%
i20	−1.61%	Nan	4.65%	−25.00%	0.28%	Nan	1.11%	−8.33%

Table 5. Summary of Performance by Number of Protocols (*n_prot*)

n_prot	MAWI		CAIDA		Unicauca		Averages	
	Prot RP	Seq RP	Prot RP	Seq RP	Prot RP	Seq RP	Av Prot	No Uni
6	0.8553	−0.6458	−0.1235	−0.4138	1.10%	−100%	24.76%	−52.98%
9	0.803	−0.7083	−0.0458	−0.125	9.12%	600%	28.28%	−41.67%
12	5.93%	37.50%	4.59%	25.00%	11.36%	Nan	7.29%	31.25%

Phi-2 with an **interval of** $i = 5$, $Seq = 2$, **and** $n\_prot = 12$

AIRBUS

Experiment Results

Experimental methodology step-by-step

01

Comparison Study of LLMs for Protocol Sequence Prediction

02

Offline Evaluation for Retrospective Drift Analysis

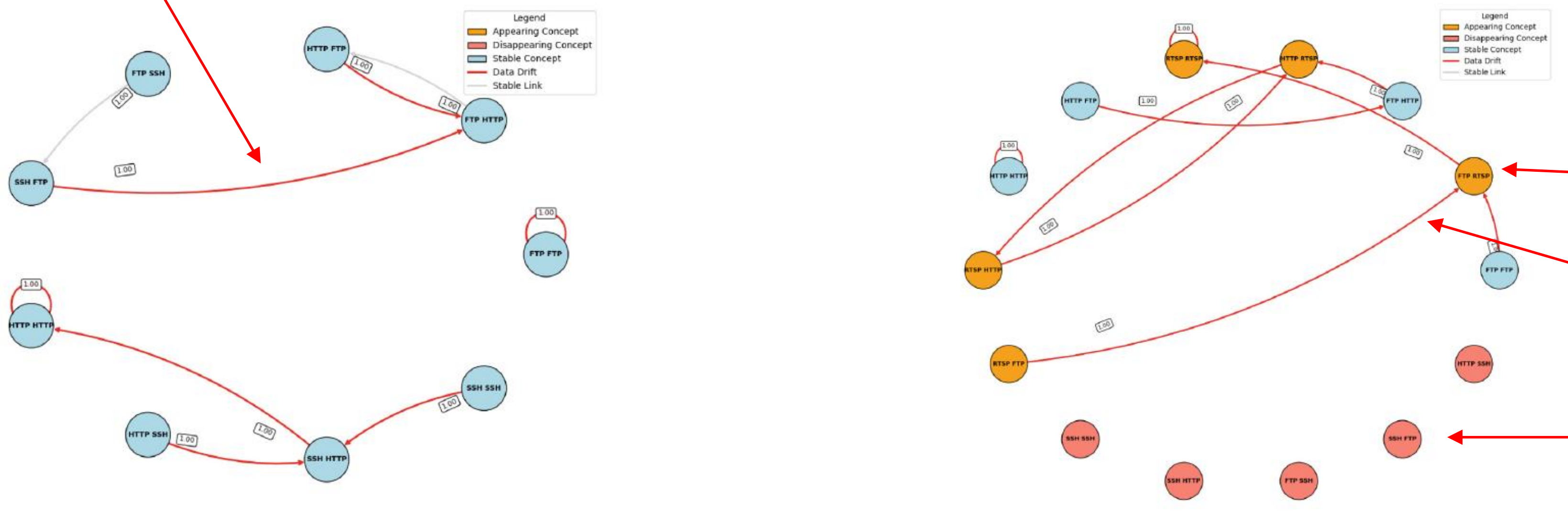
03

Online Evaluation for Real-time Drift Analysis

Experiment Results

Offline Evaluation for Retrospective Drift Analysis

Data drift between two
well known concepts



(a) Data Drift Set

(b) Concept and Data Drift Set

Fig. 1. Interpretability of Model Drift: Evidence from Simulated Data Patterns

Appearing
concept

Suspicious Data Link ➔
flagged as Data drift

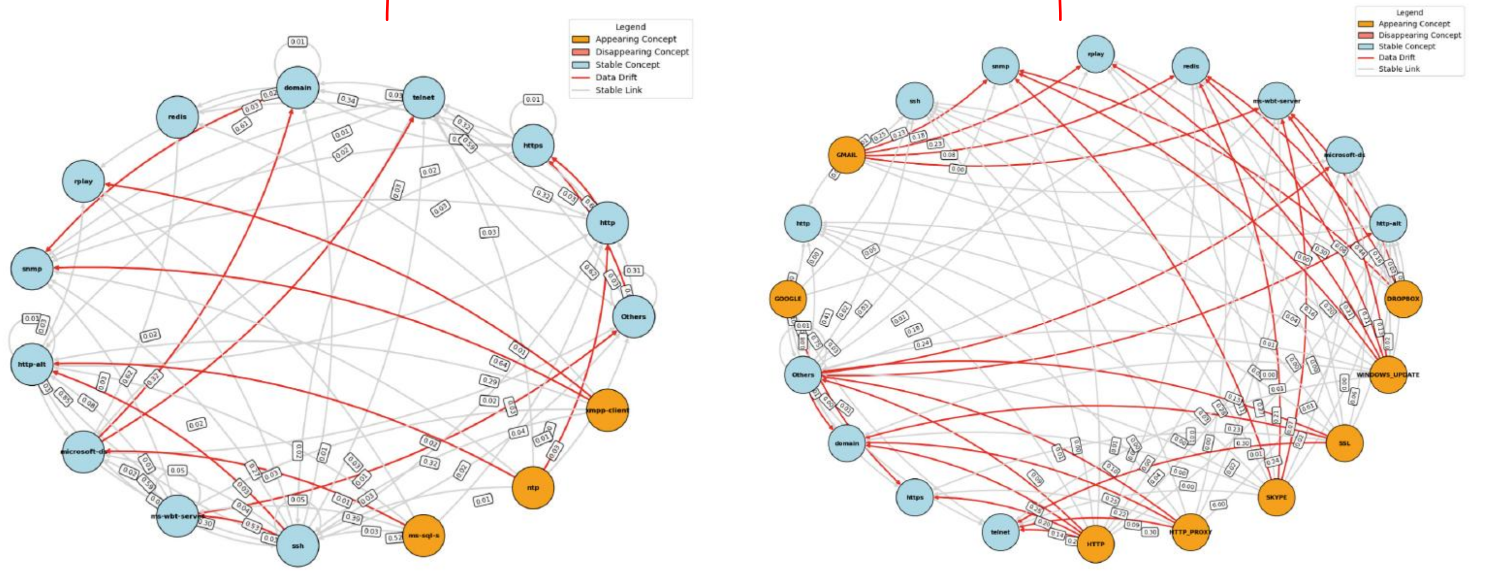
Disappearing
concept

AIRBUS

Experiment Results

Offline Evaluation for Retrospective Drift Analysis

Visualization complexity management ➔ Top k Node



(a) Caida (b) Unicauca

Fig. 2. Interpretability of Model Drift: Validation on MAWI Data Patterns

AIRBUS

Experiment Results

Experimental methodology step-by-step

01

Comparison Study of LLMs for Protocol Sequence Prediction

02

Offline Evaluation for Retrospective Drift Analysis

03

Online Evaluation for Real-time Drift Analysis

Experiment Results

Online Evaluation for Real-time Drift Analysis

Table 6. Top 15 Results of Grid Search for Drift Detection Configurations

Step	Win.	Metric	Bias	Drift	Mode	Sens.	Buf.	Thresh.	Accuracy	AUC
100	300	CE	Mean	Mean	Semi	3.0	—	—	0.985	1.000
100	300	CE	Mean	Mean	Semi	4.0	—	—	0.985	1.000
100	300	CE	Mean	Mean	Full	—	0.1	—	0.985	1.000
100	300	CE	Max	Mean	Semi	3.0	—	—	0.985	1.000
100	300	CE	Max	Mean	Semi	4.0	—	—	0.985	1.000
100	300	CE	Max	Mean	Full	—	0.1	—	0.985	1.000
100	300	CE	Mean	Max	No	—	—	3	0.985	0.933
100	300	CE	Max	Max	No	—	—	3	0.985	0.933
50	300	CE	Mean	Mean	Semi	3.0	—	—	0.982	0.999
50	300	CE	Mean	Mean	Semi	4.0	—	—	0.982	0.999
50	300	CE	Mean	Mean	Full	—	0.1	—	0.982	0.999
50	300	CE	Max	Mean	Semi	3.0	—	—	0.982	0.999
50	300	CE	Max	Mean	Semi	4.0	—	—	0.982	0.999
50	300	CE	Max	Mean	Full	—	0.1	—	0.982	0.999
50	300	CE	Mean	Max	No	—	—	3	0.982	0.895

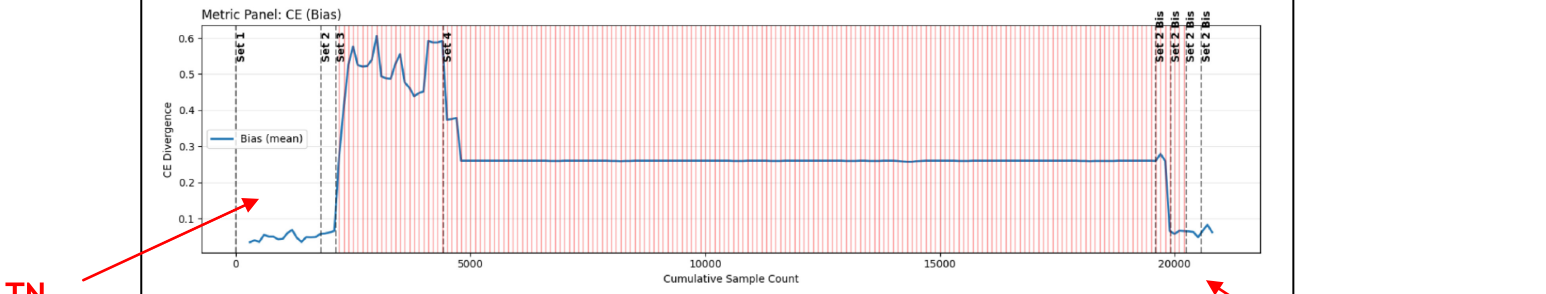
Best

Average

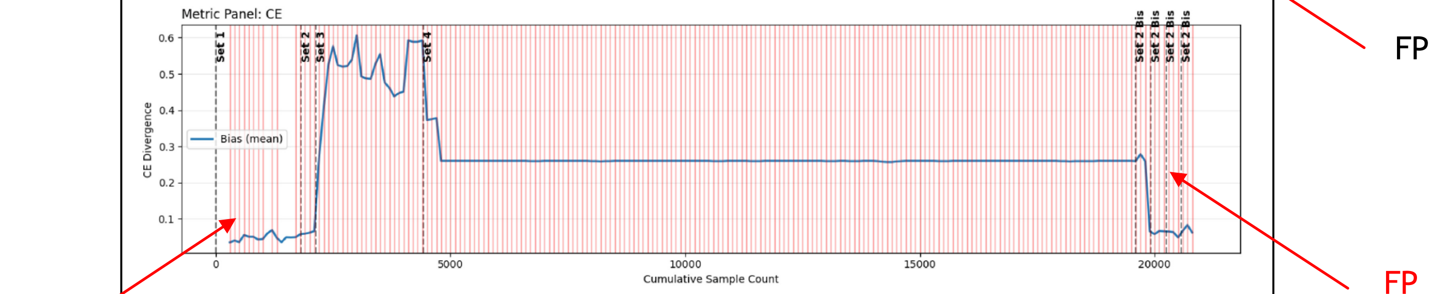
AIRBUS

Experiment Results

Online Evaluation for Real-time Drift Analysis



(a) Best config : {100, 300, "ce", "mean" , "mean", "full", 0.1}



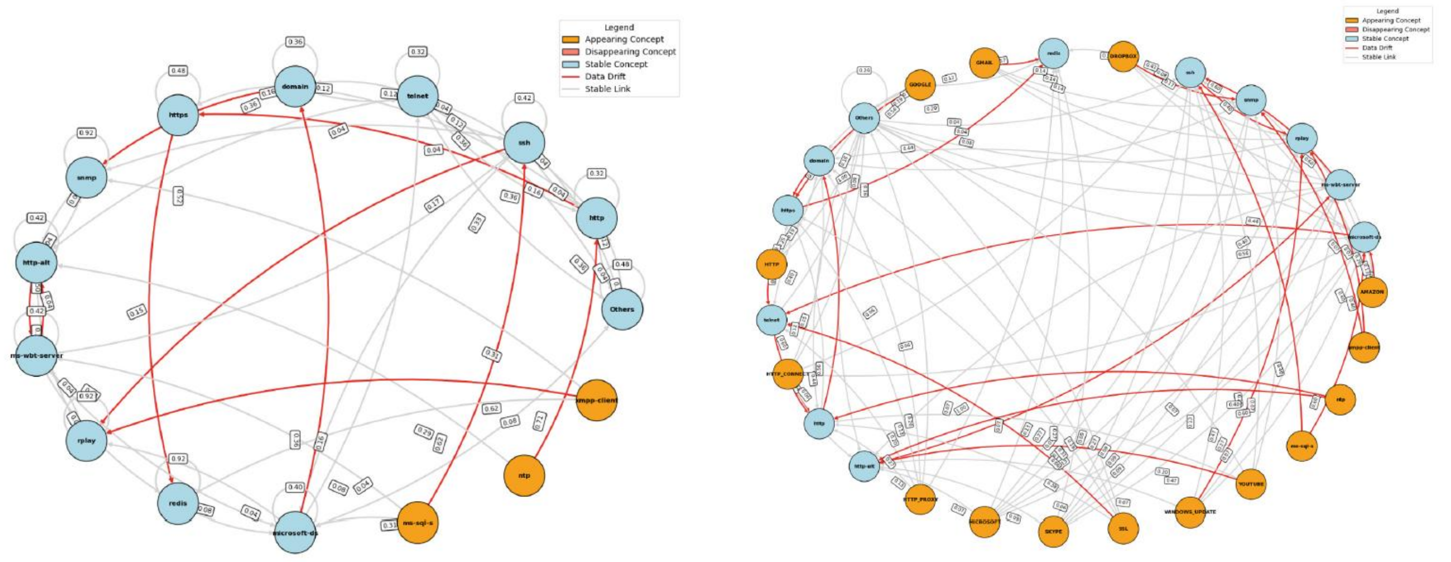
(b) Average config : {100, 300, "ce", "mean" , "max", "no", 1}

Fig. 3. Online Detection of Model Drift: Best and Average Config for MAWI Data Patterns

Experiment Results

Online Evaluation for Real-time Drift Analysis

<u>Still accurate in online detection and eXplanation (≈ 300 online observations)</u>
even if not the full dataset



(a) 1st Drift Detected on Set 3 Caida (b) 1st Drift Detected on Set 4 Unicauca

Fig. 4. Online detection and eXplanation of Model Drift trained on Mawi

Conclusion

- <u>Feasibility Study</u> ➔ **Tokenizer and LLM for Network Protocols Prediction**
  - Universal Representation
  - Transfer Learning

Relative performance shows the added value in real world scenarios of network protocol prediction.

- <u>Proof of Concept</u> ➔ **XD model-agnostic framework**
  - Pattern Modeling for Autoregressive Models;
  - Unsupervised thresholding mechanism for Drift Detection;
  - Concept and Data Drift Visual eXplanation.

Results show that **XD** successfully bridges the gap **between** **<u>high-dimensional statistical shifts</u>** **and** **<u>user-centric topological explanations</u>**.

- On-going works
  - Extend framework for continues-value time series, and extend to broader NLP tasks.

AIRBUS

Thank You !

Contact:

- Professional mail : abhishek.djeachandrane@airbus.com
- LinkedIn : www.linkedin.com/in/abhishek-djeachandrane

- Repository (Publicly available) :

  https://gitlab.ailab.airbus.com/abhishek/xd4llm