EXPLAINABLE ANOMALY DETECTION IN NETWORK TRAFFIC USING LLM

Kamil Jeřábek, *CESNET a.l.e. & BUT*

***Josef Koumar**,* CESNET a.l.e. & CTU in Prague

Jiří Setinský, *CESNET a.l.e. & BUT*

Jaroslav Pešek, *CESNET a.l.e. & CTU in Prague*

Paper accepted to The 38th IEEE/IFIP Network Operations and Management Symposium

MOTIVATION

- Rising of encrypted traffic ==> statistical anomaly detection
- Operators face hundreds of daily anomalies, most of them lacking any clear explanation. This leads to:
  - Analyst fatigue
  - Missed incidents
  - Inefficient resource allocation
- Current tools detect, but don’t explain
- What if LLMs could act as a virtual analyst?

OUR APPROACH

Detected anomalies are not explainable for SoC
==> Let the LLM explain these alerts

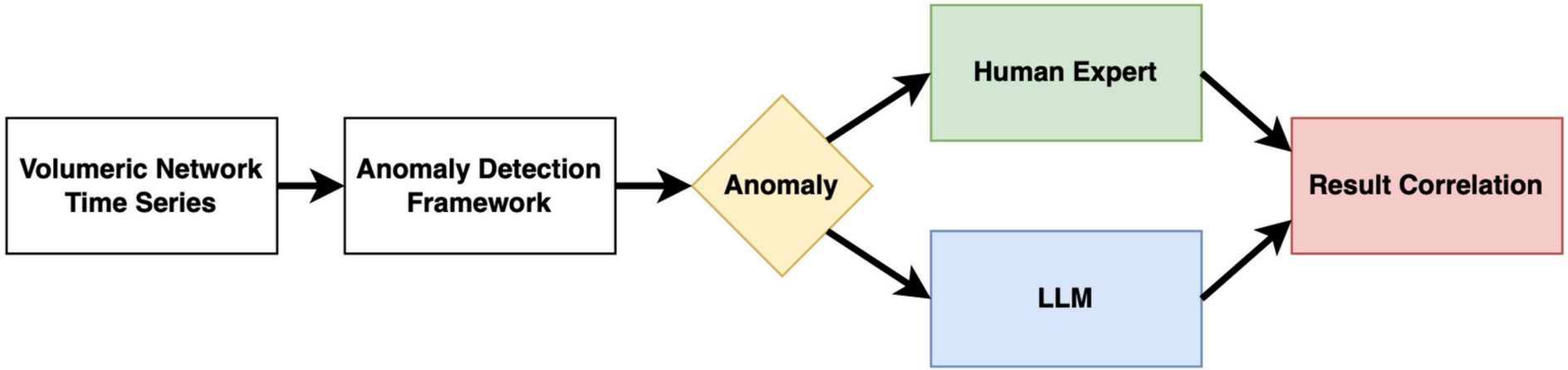
This setup:

- reduces the need for heavy computation because the LLM is only used when an anomaly has already been identified
- is zeroshot, meaning the LLM relies on its internal, general knowledge without extra training on our specific data

To demonstrate how this works, we test three different case studies

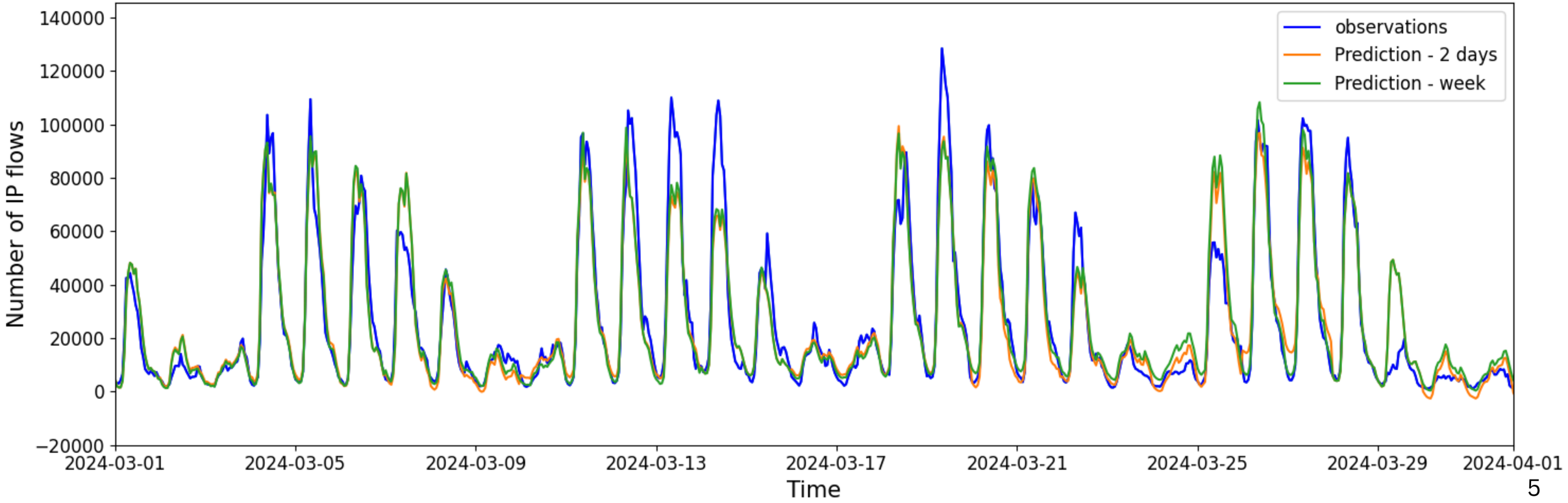
METHODOLOGY

1. Anomaly detection using network traffic forecasting
2. Annotation of anomalies as human expert
3. Prompt tuning
4. Prompting
5. Evaluation



ANOMALY DETECTION

1. Creation of time series from network traffic
2. Apply time series forecasting
3. Comparison of observation with predicted values



DATASET

We created **CESNET TimeSeries24 dataset** which contains time series created from **66 billion IP flows** that contain **4 trillion packets** that carry approximately **3.7 petabytes of data**

Time Series Metrics:

- Number of IP flows, packets, bytes
- Number of unique destination IP addresses
- Number of unique destination ASNs
- Number of unique destination countries
- TCP/UDP ratio
- Packet direction ratio
- Average TTL and duration of IP flows

Aggegation:

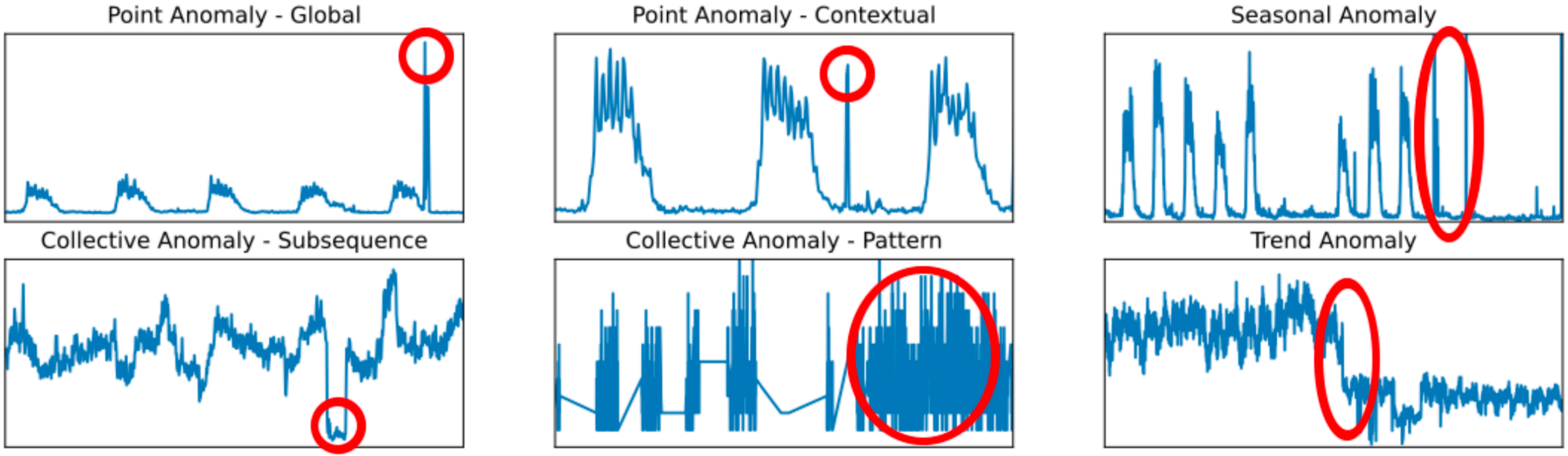
- 10 minutes
- 1 hour
- 1 day

Identifiers:

- IP addresses
- Institutions
- Institution subnets

ANNOTATION

- Lack of anomaly labels in the data
- Pick anomalies manually by experts
- Labeling based on 3 experts agreement
- 70 annotated anomalies
  - 17 for prompt tuning
  - 53 held out for evaluation



PROMPT TUNNING

- No model fine-tuning (zero shot)
  - Using general existing models with prompt tuning only

- Prompt
  - Context about the data
    - how anomalies was detected
  - Surrounding values before and after
    - 48 hours and 7 days time window mean and average, anomaly detected, each metric
  - Time of occurrence
  - Output format specification

EVALUATION PROCEDURE

- 5 available general models
  - GPT-4, GPT-4o, GPT-4o-mini, Gemini 2 flash experimental
- Assessed against human experts annotation
  - Assigned priority: zero, low, medium, high, and critical
  - Set threshold for model outcomes
- F1 and Precision metric

Materials:

CASE 1: LLM AS AUTONOMOUS ANOMALY ANALYST AGENT

- **Decision:** Threat or false positive
- **Context:** past and future
  - Info from 48 hours 7 days before and after incident (mean, std)

Model	F1	Precision	Priority
GPT-4	0.81	0.92	$> medium$
GPT-4o	0.86	0.92	$> medium$
GPT-4o-mini	0.75	0.81	$>= medium$
Gemini 2.0 Flash Experimental	0.92	0.93	$> medium$

CASE 2: LLM AS AUTONOMOUS ANOMALY ALERTER AGENT

- Simulates real time anomaly detection
- **Decision:** Threat or false positive
- **Context:** Before incident only
  - Infor from 48 hours and 7 days before (mean, std)

Model	F1	Precision	Priority
GPT-4	0.92	0.88	$>= medium$
GPT-4o	0.89	0.96	$> medium$
GPT-4o-mini	0.81	0.83	$>= medium$
Gemini 2.0 Flash Experimental	0.93	0.96	$> medium$

CASE 3: LLM AS A HUMAN ANALYST COMPANION

- **GPT-4**
  - Mostly accurate attack identifications
  - Occasionally underestimated severity
- **GPT-4o**
  - Produced the most detailed and actionable expla-nations
  - Closely matching human expert assessments
- **Gemini 2.0 Flash Experimental**
  - Performed well
  - Close to GPT-4o in explaining anomalous behavior
- **GPT-4o-mini**
  - While capable of describing network patterns
  - Struggled with deeper insights and concrete attack attributions

CONCLUSION

- We integrate LLM with an anomaly detection framework to enhance explainability in network traffic analysis
- Our approach demonstrates LLMs potential in false positive reduction, autonomous decision-making and improved explainability for network analysts
- We evaluate our method on real-world data against three human experts annotation agreement, demonstrating its effectiveness in cybersecurity applications

cesnet

BRNO UNIVERSITY OF TECHNOLOGY

THANK YOU FOR YOUR ATTENTION

cesnet

BRNO UNIVERSITY OF TECHNOLOGY