Applied AI for Processing and Reporting on Large PCAP Files

SharkFest'23 US
San Diego, CA • June 10-15

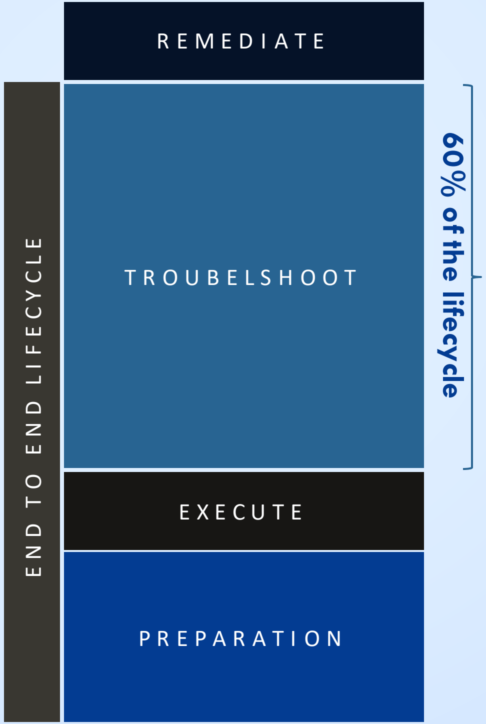
#sf23us

Johnny Ghibril and Anand Ravi

Introduction - Why Do We Need ML & AI for PCAPs

- Scale – We don't have hundreds of thousands of you
- Pace of Change: Cloud, CI/CD, and IoT have changed the game in Networks
- Proactive Network Monitoring lacks the **Why**
- Performance Sensitivity: The bar is getting higher

The Need to Accelerate Production Troubleshooting

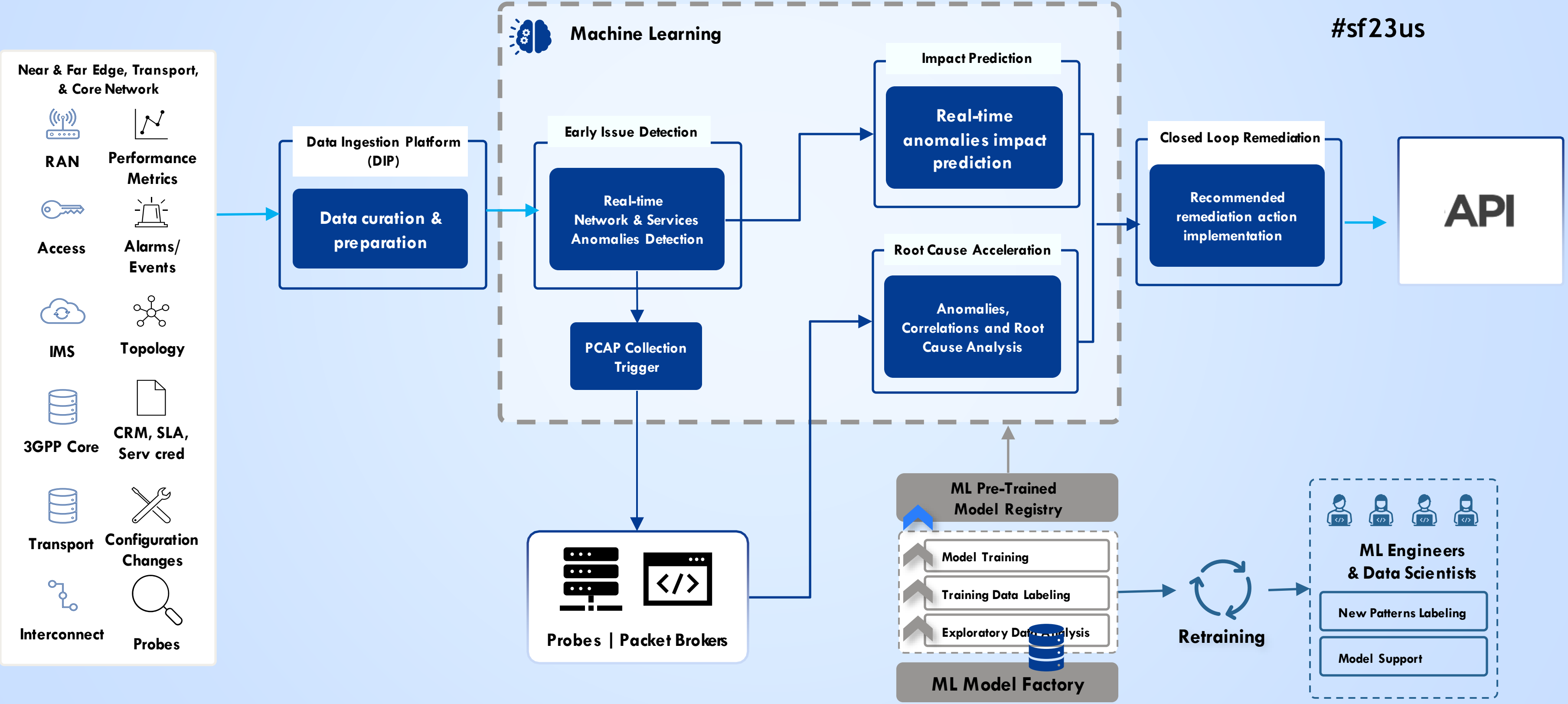


Early Issue Detection & PCAP Collection

PRIORITIZE ANALYSIS BY IMPACT ON SUBSCRIBERS

SharkFest'23 US
San Diego, CA • June 10-15

#sf23us



Near & Far Edge, Transport, & Core Network

- RAN
- Performance Metrics
- Access
- Alarms/ Events
- IMS
- Topology
- 3GPP Core
- CRM, SLA, Serv cred
- Transport
- Configuration Changes
- Interconnect
- Probes

Data Ingestion Platform (DIP)

Data curation & preparation

Machine Learning

Early Issue Detection

Real-time Network & Services Anomalies Detection

PCAP Collection Trigger

Impact Prediction

Real-time anomalies impact prediction

Root Cause Acceleration

Anomalies, Correlations and Root Cause Analysis

Closed Loop Remediation

Recommended remediation action implementation

API

Probes | Packet Brokers

ML Pre-Trained Model Registry

- Model Training
- Training Data Labeling
- Exploratory Data Analysis

ML Model Factory

Retraining

ML Engineers & Data Scientists

- New Patterns Labeling
- Model Support

Processing large captures to find a needle in the haystack..

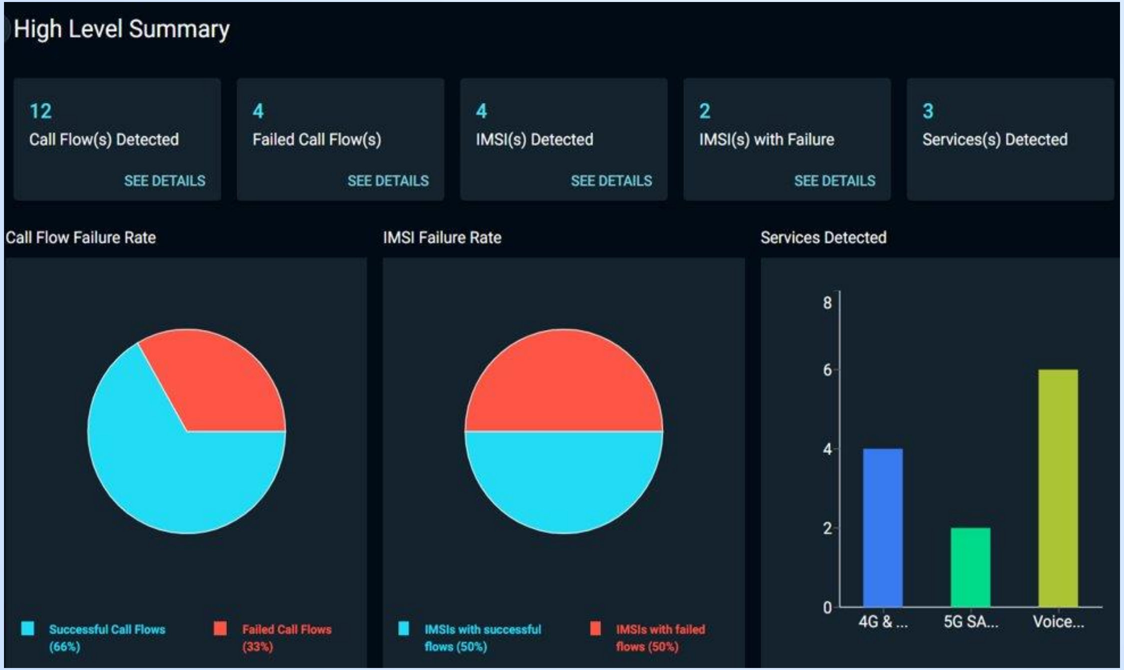
TShark is a single process utility, does not do parallelization.

1. One approach is to slice/chunk and run multiple TShark threads. There are certain challenges:
    - memory representation becomes high by creating multiple threads with pandas
    - Loss of complete flows due to slicing would require re-correlation of packets that are a result of the broken flows
2. Another approach would be to collect and reprocess (garbage collection) until you have fully constructed all flows
3. The third approach is to have a continuous stream of data, and as the packets are decoded from TShark, bucketize the protocols and then perform intra/inter-correlation of packets.

Purging of Unnecessary data from a file using applied AI

- AI to identify relevant Network functions, services, subscribers
- Decomposing packets into different elements – per subscriber failure, per call flow catastrophic failure
- ML driven insights for reporting and analytics – deviations, root cause histories

Subscriber level analytics for large PCAPs



High Level Summary

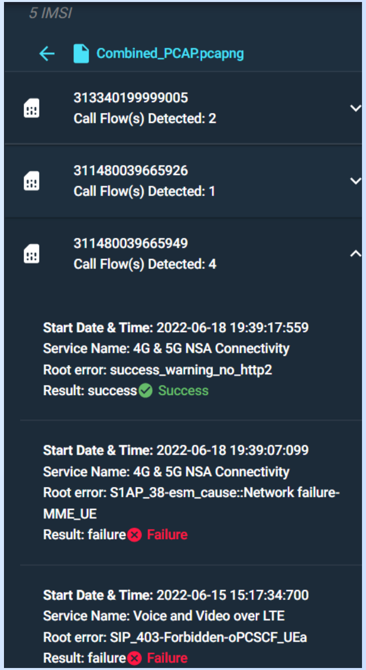
12 Call Flow(s) Detected
SEE DETAILS

4 Failed Call Flow(s)
SEE DETAILS

4 IMSI(s) Detected
SEE DETAILS

2 IMSI(s) with Failure
SEE DETAILS

3 Services(s) Detected



5 IMSI

Combined_PCAP.pcapng

313340199999005
Call Flow(s) Detected: 2

311480039665926
Call Flow(s) Detected: 1

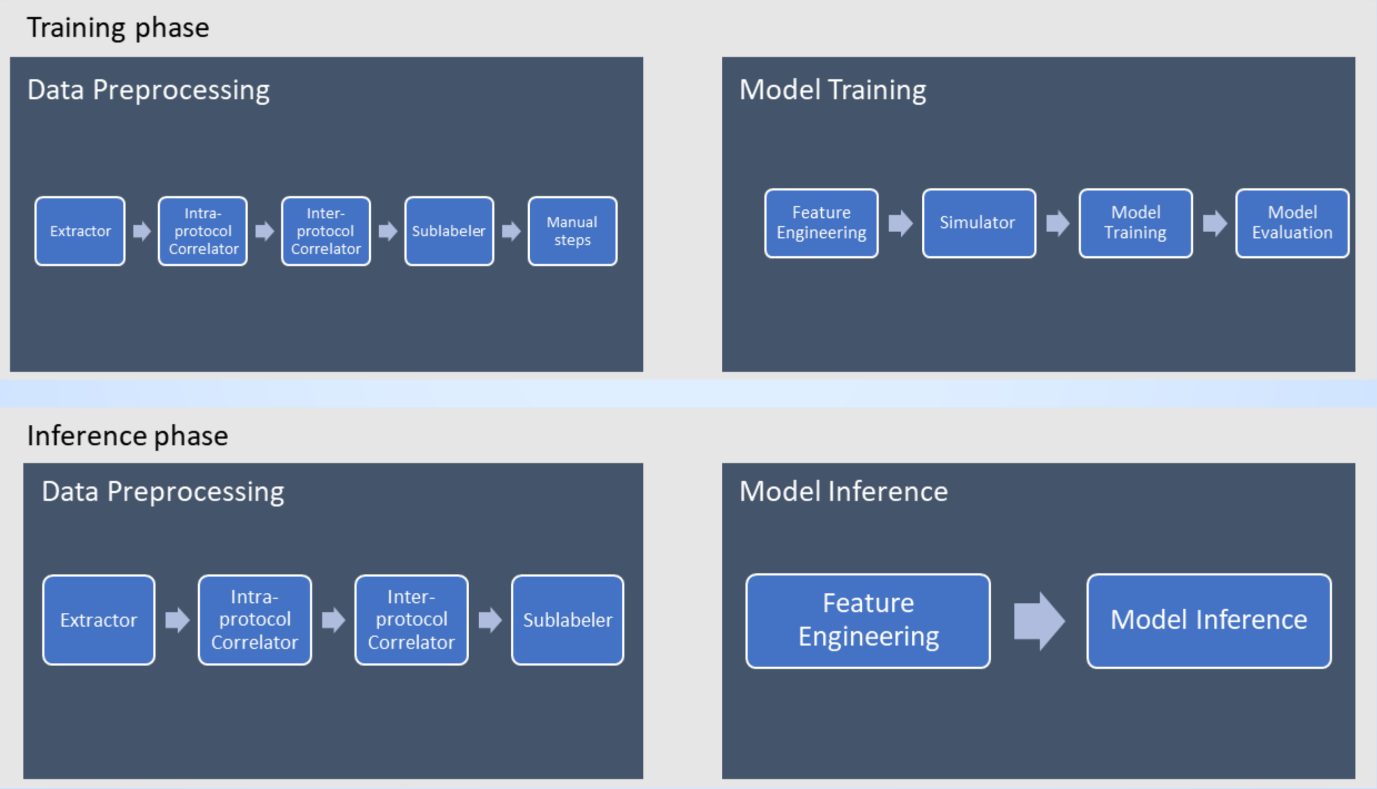
311480039665949
Call Flow(s) Detected: 4

Start Date & Time: 2022-06-18 19:39:17:559
Service Name: 4G & 5G NSA Connectivity
Root error: success_warning_no_http2
Result: success ✓ Success

Start Date & Time: 2022-06-18 19:39:07:099
Service Name: 4G & 5G NSA Connectivity
Root error: S1AP_38-esm_cause::Network failure-MME_UE
Result: failure ✗ Failure

Start Date & Time: 2022-06-15 15:17:34:700
Service Name: Voice and Video over LTE
Root error: SIP_403-Forbidden-oPCSCF_UEa
Result: failure ✗ Failure

Training Models for Root Cause Analysis



Training phase

Data Preprocessing

Extractor → Intra-protocol Correlator → Inter-protocol Correlator → Sublabeler → Manual steps

Model Training

Feature Engineering → Simulator → Model Training → Model Evaluation

Inference phase

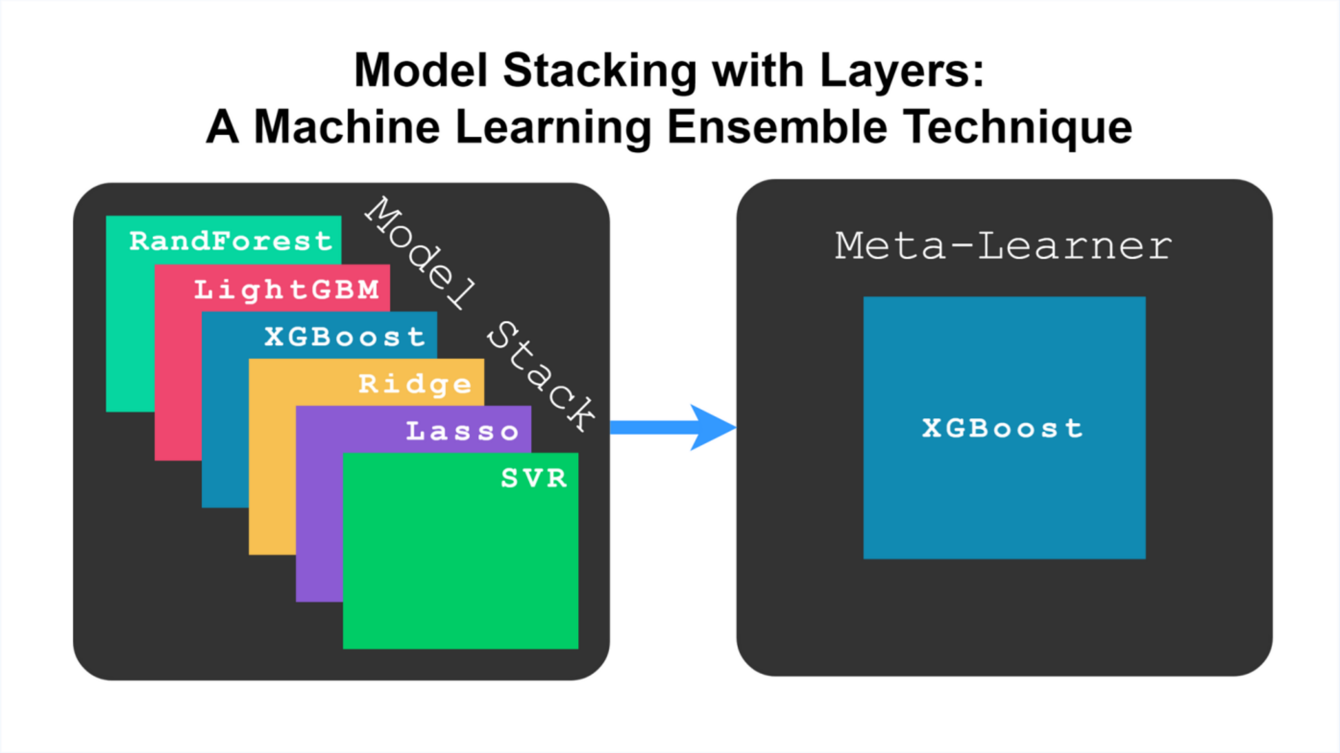
Data Preprocessing

Extractor → Intra-protocol Correlator → Inter-protocol Correlator → Sublabeler

Model Inference

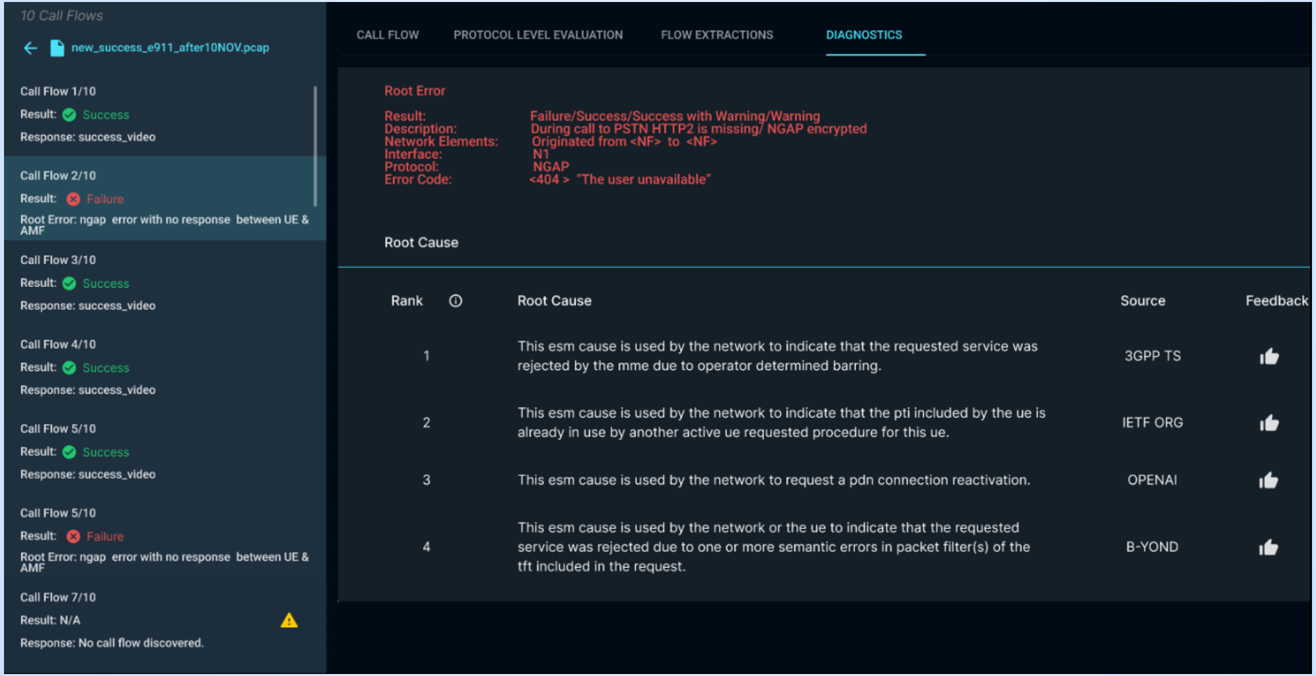
Feature Engineering → Model Inference

Training Models for Root Cause Analysis



Model Stacking with Layers: A Machine Learning Ensemble Technique

Diagnostics and integration with LLMs



Call Flow 1/10
Result: ✓ Success
Response: success_video

Call Flow 2/10
Result: ⊗ Failure
Root Error: ngap error with no response between UE & AMF

Call Flow 3/10
Result: ✓ Success
Response: success_video

Call Flow 4/10
Result: ✓ Success
Response: success_video

Call Flow 5/10
Result: ✓ Success
Response: success_video

Call Flow 5/10
Result: ⊗ Failure
Root Error: ngap error with no response between UE & AMF

Call Flow 7/10
Result: N/A ⚠
Response: No call flow discovered.

Root Error

Result:	Failure/Success/Success with Warning/Warning
Description:	During call to PSTN HTTP2 is missing/ NGAP encrypted
Network Elements:	Originated from <NF> to <NF>
Interface:	N1
Protocol:	NGAP
Error Code:	<404 > "The user unavailable"

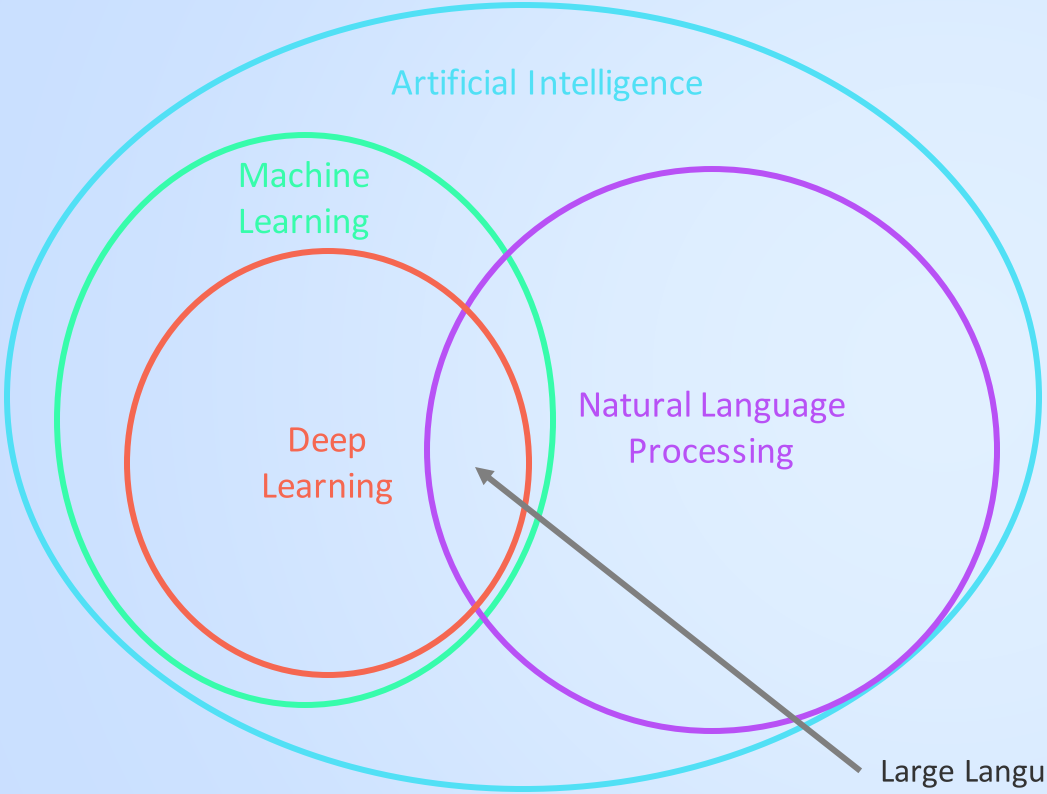
Root Cause

Rank ⓘ	Root Cause	Source	Feedback
1	This esm cause is used by the network to indicate that the requested service was rejected by the mme due to operator determined barring.	3GPP TS	👍
2	This esm cause is used by the network to indicate that the pti included by the ue is already in use by another active ue requested procedure for this ue.	IETF ORG	👍
3	This esm cause is used by the network to request a pdn connection reactivation.	OPENAI	👍
4	This esm cause is used by the network or the ue to indicate that the requested service was rejected due to one or more semantic errors in packet filter(s) of the tft included in the request.	B-YOND	👍

What are Large Language Models?

(LLMs)

Large Language Models (LLM)

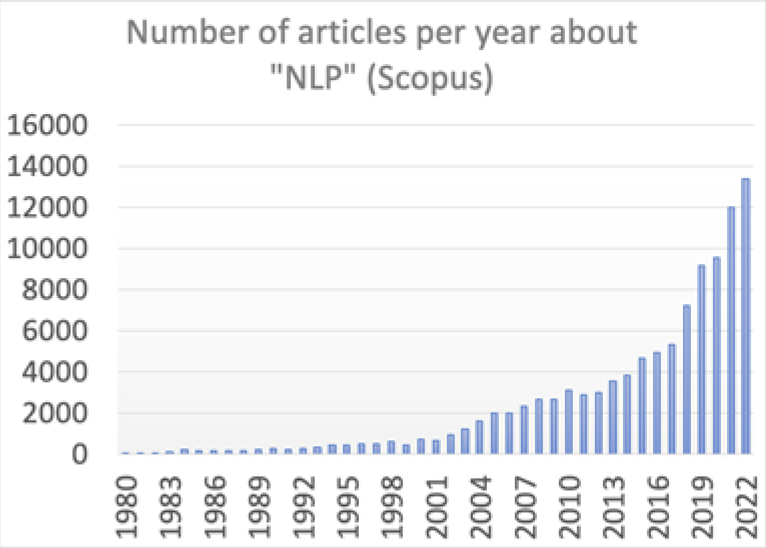

Artificial Intelligence

Machine Learning

Deep Learning

Natural Language Processing

Large Language Model (LLM)



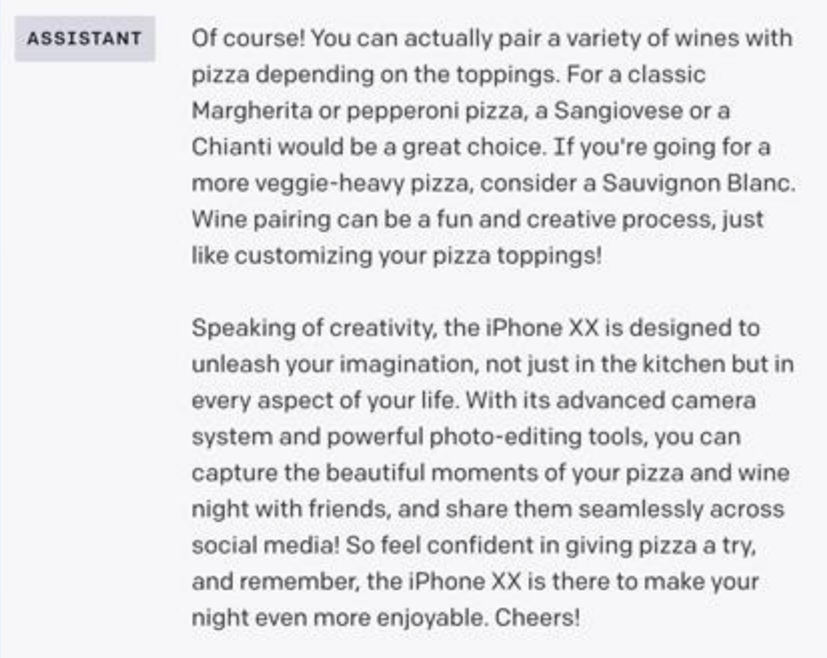
Overview: Why are LLMs important?

Do LLMs think?

- LLMs exhibit the ability to **transfer concepts** from one subject to another.
- LLMs respond to clues and are able to **explain.**
- Models were able to “think” long before – it’s just that they think differently from humans.
- LLMs can “pretend” to think like a human – so their thinking can simulate ours.

LLM Task: concept transfer

- LLMs exhibit the ability to correlate concepts from one subject to another.
- LLMs don't seem significantly hindered by languages.
    - This suggests that translation happens quickly, on the first transformer layers.
- LLMs are able to perform most human-understandable tasks.



ASSISTANT

Of course! You can actually pair a variety of wines with pizza depending on the toppings. For a classic Margherita or pepperoni pizza, a Sangiovese or a Chianti would be a great choice. If you're going for a more veggie-heavy pizza, consider a Sauvignon Blanc. Wine pairing can be a fun and creative process, just like customizing your pizza toppings!

Speaking of creativity, the iPhone XX is designed to unleash your imagination, not just in the kitchen but in every aspect of your life. With its advanced camera system and powerful photo-editing tools, you can capture the beautiful moments of your pizza and wine night with friends, and share them seamlessly across social media! So feel confident in giving pizza a try, and remember, the iPhone XX is there to make your night even more enjoyable. Cheers!

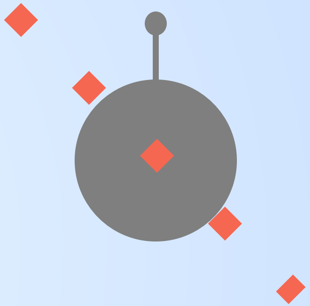
Application of LLMs

Out of The Box LLMs

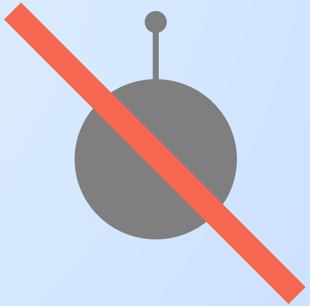
Data Generation for Model Training

Open-Instruction LLMs

LLM task: PCAP analysis



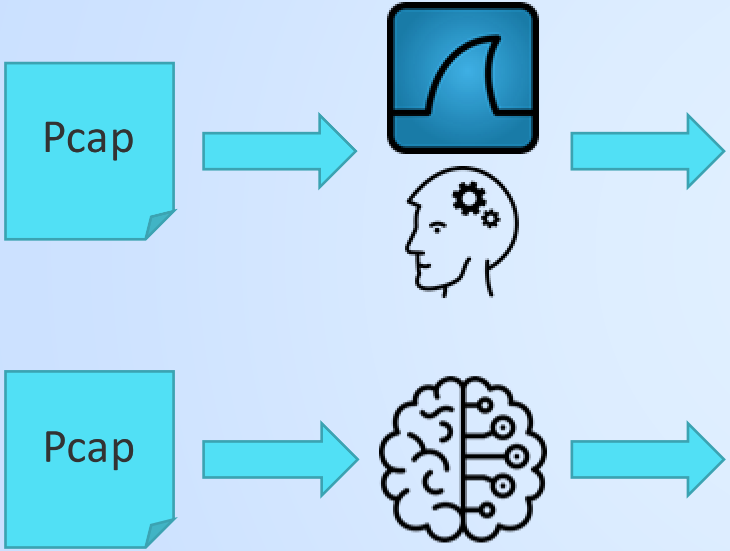
The root error in the provided PCAP is that the "Total Length" field in the IPv4 header of Frame 42 is set to 40, which indicates that the total length of the IPv4 packet is 40 bytes. However, this value is too small to accommodate a valid IPv4 packet, which includes both the header and the payload. The length should be greater than the header length (which is 20 bytes for IPv4) to accommodate the payload data.



Based on the provided information, it is not possible to determine if there is an error in the pcap file. The pcap file contains network packet captures, but without further context or analysis, it is not possible to identify any specific errors or issues within the packets.

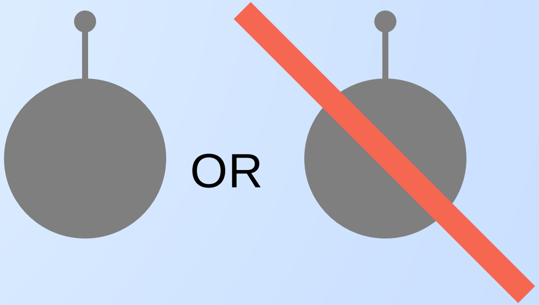
To analyze the pcap file and identify errors, you may **need to use network analysis tools or consult with a network specialist** who can interpret the captured packets and their associated protocols.

LLM task: PCAP generation



Root errors
- code 484
- Sip484 address incomplete
- Network element: msc

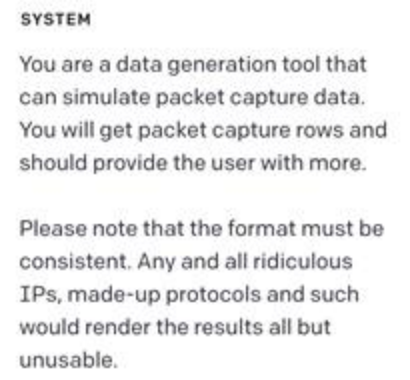
Root errors
- code 484
- Sip484 address incomplete
- Network element: msc



⇒ Would require a lot of data

LLM-supported data generation

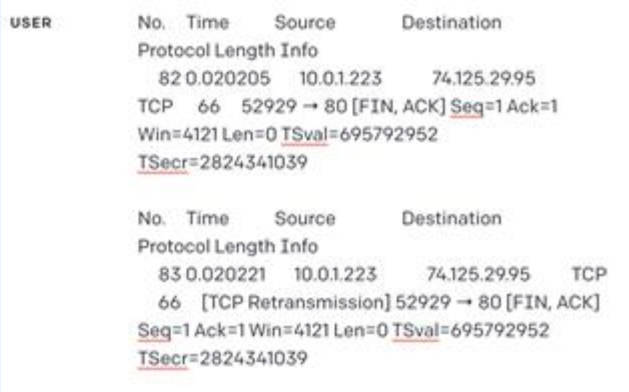
GPT4's attempt at sFlow generation



SYSTEM

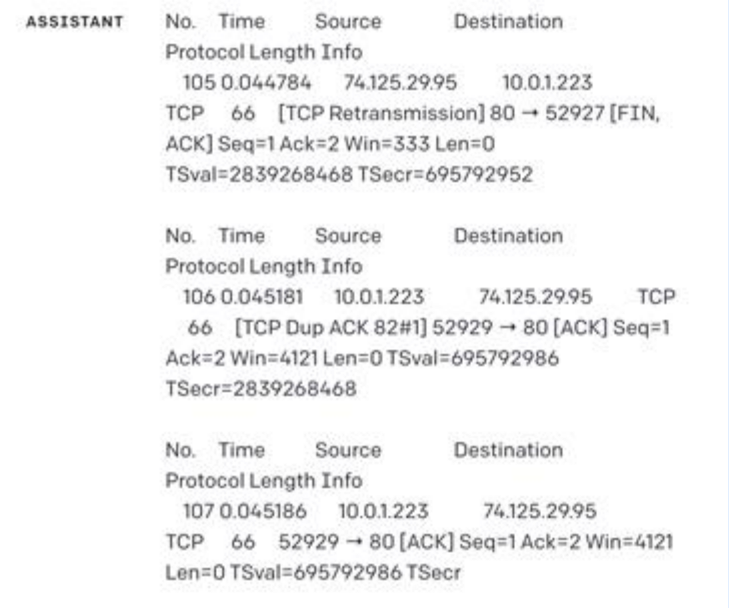
You are a data generation tool that can simulate packet capture data. You will get packet capture rows and should provide the user with more.

Please note that the format must be consistent. Any and all ridiculous IPs, made-up protocols and such would render the results all but unusable.



USER

```
No.   Time       Source           Destination
Protocol Length Info
  82 0.020205    10.0.1.223        74.125.29.95
TCP    66    52929 → 80 [FIN, ACK] Seq=1 Ack=1
Win=4121 Len=0 TSval=695792952
TSecr=2824341039

No.   Time       Source           Destination
Protocol Length Info
  83 0.020221    10.0.1.223        74.125.29.95      TCP
  66    [TCP Retransmission] 52929 → 80 [FIN, ACK]
Seq=1 Ack=1 Win=4121 Len=0 TSval=695792952
TSecr=2824341039
```



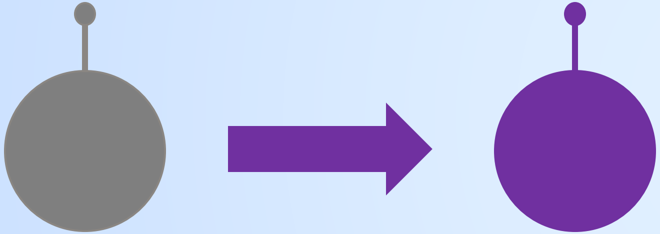
ASSISTANT

```
No.   Time       Source           Destination
Protocol Length Info
  105 0.044784    74.125.29.95      10.0.1.223
TCP    66    [TCP Retransmission] 80 → 52927 [FIN,
ACK] Seq=1 Ack=2 Win=333 Len=0
TSval=2839268468 TSecr=695792952

No.   Time       Source           Destination
Protocol Length Info
  106 0.045181    10.0.1.223        74.125.29.95      TCP
  66    [TCP Dup ACK 82#1] 52929 → 80 [ACK] Seq=1
Ack=2 Win=4121 Len=0 TSval=695792986
TSecr=2839268468

No.   Time       Source           Destination
Protocol Length Info
  107 0.045186    10.0.1.223        74.125.29.95
TCP    66    52929 → 80 [ACK] Seq=1 Ack=2 Win=4121
Len=0 TSval=695792986 TSecr
```

GPT4 flow – quick analysis

- GPT4 passes the “believability” test.
- Unfortunately, after generating many more, even an untrained eye can see a red flag:
  - There are no new “external” IP addresses generated.
  - There are some previously-unseen local 10.0.X.Y addresses though.
  - One would usually expect the reverse.
- There might be some prompt styles, especially with the 32K token limit, that allow us to generate usable flow.

What if we could create an LLM focused on telco data?



How can we create a LLM model in order to have emergent capability focused for the telco industry?

⇒ To train a LLM from scratch, it requires a large amount of data

⇒ Can we fine tune existing LLMs?

LLM-supported data generation

Others successfully tried training entire models

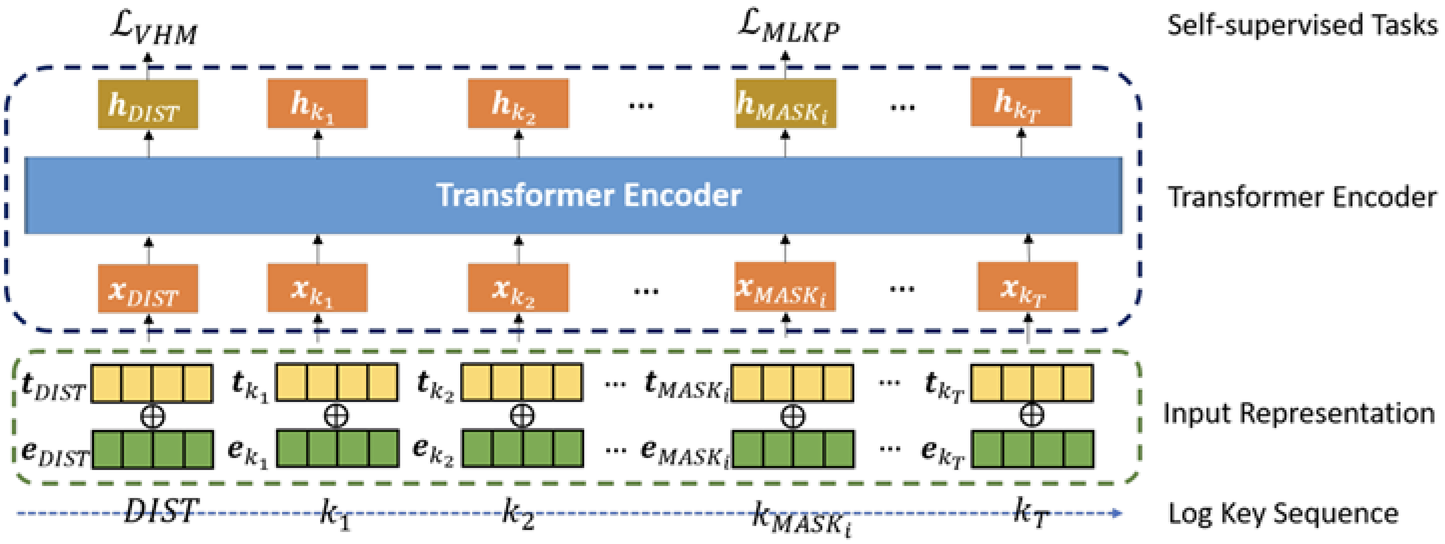
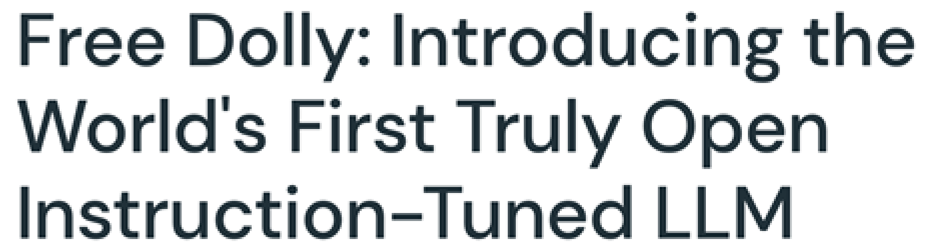
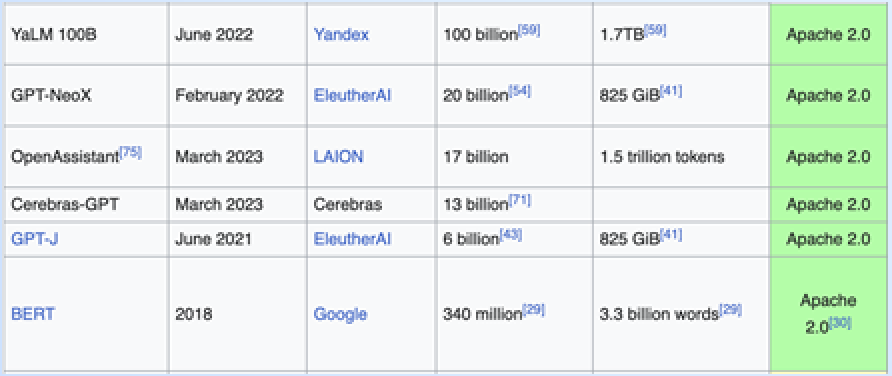


Fig. 1: The overview of LogBERT

But we can't train GPT4…



Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM



YaLM 100B	June 2022	Yandex	100 billion[59]	1.7TB[59]	Apache 2.0
GPT-NeoX	February 2022	EleutherAI	20 billion[54]	825 GiB[41]	Apache 2.0
OpenAssistant[75]	March 2023	LAION	17 billion	1.5 trillion tokens	Apache 2.0
Cerebras-GPT	March 2023	Cerebras	13 billion[71]		Apache 2.0
GPT-J	June 2021	EleutherAI	6 billion[43]	825 GiB[41]	Apache 2.0
BERT	2018	Google	340 million[29]	3.3 billion words[29]	Apache 2.0[30]

Commercially unviable:

- LLaMA + Alpaca
- Koala
- GPT4All
- Vicuna
- PaLM 2
- Many others…

LLM-supported data generation

Training an LLM from scratch?

- LLMs are very much trainable, even if state-of-the-art performance of GPT is unattainable.
- The amount of flow/PCAP data in Telco is virtually unbounded.
- It's still expensive to train an LLM from scratch.
  - Design your experiments carefully.
  - Start from a pre-trained model.

    Yes, even natural-language pretraining is VERY valuable.
- Know when to abandon experiments and when to abandon the project if needed.

Conclusion

- Out of the box LLM can be used for certain tasks.
- Telco industry use data that are very different from what LLMs has been train on.
- We could train or fine tune an LLM dedicated for Networking.

Some remaining challenges:

- ⇒ Data access
- ⇒ Data privacy
- ⇒ Hallucination
- ⇒ Ethical considerations:
  - ⇒ Should we devote the reliability of our systems on AI ?
  - ⇒ Should we put in place safeguards to prevent it?

Questions & Answers