TelecomTS: A Multi-Modal Observability Dataset for Time Series and Language Analysis

Austin Feng*, Andreas Varvarigos*, Ioannis Panitsas, Daniela Fernandez, Jinbiao Wei, Yuwei Guo, Jialin Chen, Ali Maatouk, Leandros Tassiulas, Rex Ying

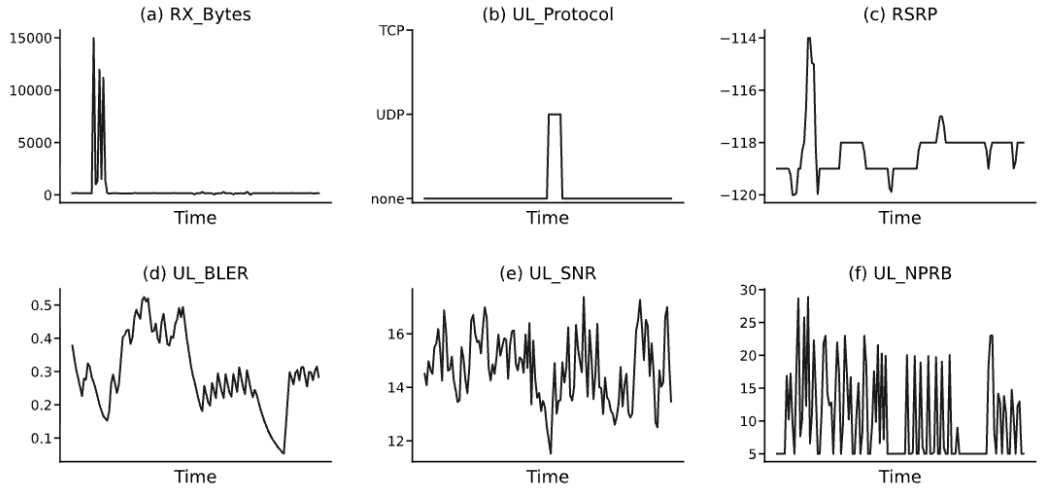
Yale SCHOOL OF ENGINEERING AND APPLIED SCIENCE

Motivation: What is Observability Data?

- Modern enterprises monitor complex systems through time series — **observability data**
- **Examples:** CPU usage, request latency, network throughput, error rates
- Domain is understudied in time series research
- Observability data breaks the assumptions of standard time series benchmarks
  - **Zero-inflated** — long inactivity, punctuated by bursts
  - **Highly dynamic** — frequent, abrupt transitions
  - **Highly stochastic** — irregular, minimal temporal structure
- **Existing dataset gaps:**
  - **Proprietary** — operators don't release their data
  - **Anonymized & normalized** — scale information is destroyed (BOOM)
  - **Forecasting-only** — no anomaly detection, root cause, or multi-modal tasks

Introducing TelecomTS

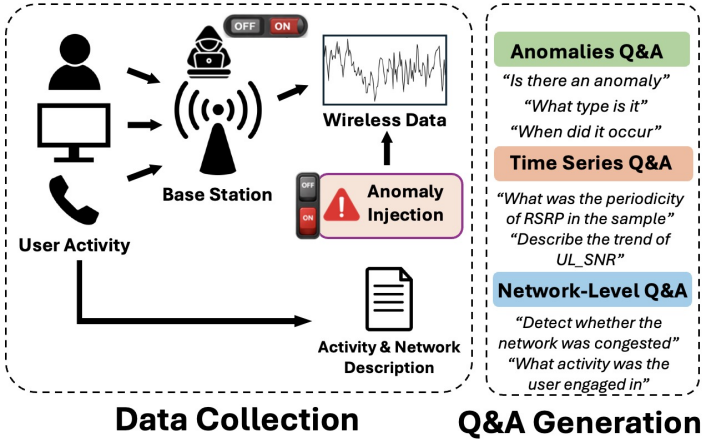
- **TelecomTS:** a large-scale 5G observability dataset built from a lab-deployed testbed
- Three contributions:
  - **First large-scale 5G observability benchmark** with heterogeneous, de-anonymized, scale-preserving covariates
  - **Multi-modal Q&A suite:** spanning anomaly, network-level, and time series Q&A tasks over real observability data
  - **Reasoning-path annotations:** explicit reasoning traces for anomaly and network-level Q&A



Q&A Suite and Reasoning Traces

- **Three Q&A families** turn the dataset into a multi-modal benchmark

1. **Anomalies Q&A** — presence, duration, root cause
   - **Real anomalies** — adversarial jammer disrupting transmissions
   - **Synthetic anomalies** — 10 literature-grounded fault types injected stochastically
2. **Network-Level Q&A** — activity, mobility, zone, congestion
3. **Time Series Q&A** — mean, variance, periodicity, trend



Example reasoning trace: Root cause identification

<think>
Okay, let's start by looking at the KPIs. The RSRP is weak but stable, which might suggest a signal strength issue, but BLER is zero, so maybe it's not a coverage problem.

▷ *Examining channel conditions*

The MCS values are fixed for UL and DL, which could indicate a lack of adaptability. The UL SNR is varying between 6.3 and 14.6 dB, which is a wide range. Even though the average is okay, the variability might imply interference.

▷ *Cross-referencing utilization and throughput*

The PRB utilization is very low in both directions, which is unexpected if the SNR is fluctuating. Since low utilization and stable BLER don't align with a coverage or capacity issue, maybe the interference is causing periods where the channel is usable but not stable. The low throughput and minimal packet activity support that the link isn't being used effectively despite the stable RSRP.

▷ *Converging on diagnosis*

This points towards interference on the same channel disrupting the signal stability. Therefore, the symptoms align with mild co-channel interference.
</think>

Answer: Co-channel interference (mild).

Anomaly Detection

- **LLMs collapse to false positives on observability data**
- GPT-4.1, o4-mini: **~1.00 recall, ~0.20 precision** — flagging everything
- Even with context, precision stays low
- **Scale-aware Mantis** (0.800 F1) and **observability-pretrained Toto** lead

Model	Precision	Recall	F1
Large language models			
GPT-4.1 (without context)	0.200	1.000	0.333
GPT-4.1 (with context)	0.173	0.609	0.270
Claude 3.7 Sonnet (without context)	0.182	0.840	0.299
Claude 3.7 Sonnet (with context)	0.194	0.860	0.322
Reasoning models			
o4-mini (without context)	0.188	1.000	0.316
o4-mini (with context)	0.246	0.580	0.345
DeepSeek-R1 (without context)	0.259	0.600	0.362
DeepSeek-R1 (with context)	0.244	0.470	0.321
Foundation Models			
Moment	0.256	0.888	0.397
Moirai2	0.346	0.490	0.405
Toto	0.521	0.750	0.615
Time series models			
Mantis	0.800	0.800	**0.800**
Mantis (w/o scaling)	0.585	0.850	0.692
TimesNet	0.389	0.652	0.487
Autoformer	0.199	0.690	0.308
Non-stat. Transformer	0.446	0.464	0.455
FEDformer	0.224	0.560	0.320
Informer	0.459	0.448	0.453
Multi-modal models			
Toto+Qwen-3-4B	0.368	0.717	0.487
Toto+Qwen-3-4B+Thinking	0.354	0.699	0.469

Time Series & Network-Level Q&A

- Multi-modal training wins on network tasks but struggles where scale matters
- **Toto+Qwen dominates network Q&A** — near-perfect on traffic, mobility, congestion
- **Lags on location** — distance estimation needs absolute signal magnitudes
- **Time series Q&A:** foundation models struggle on erratic KPIs (TX_Bytes)

Model	Time series QA					Network QA			
	Statistics		Periodicity		Trends	Traffic	Mobility	Location	Congestion
	MAE_{min}	MAE_{max}	MAE_{min}	MAE_{max}	Acc	Acc	Acc	Acc	Acc
GPT-4.1	0.163	1588.1	57.61	93.01	0.163	0.448	0.533	0.294	0.494
Claude 3.7-Sonnet	0.093	1315.8	32.04	64.04	0.109	0.414	0.950	0.428	0.461
o4-mini	0.027	247.1	37.21	63.15	0.134	0.433	0.767	0.367	0.494
DeepSeek-R1	0.020	1542.6	50.33	61.73	0.134	0.357	0.983	0.339	0.483
Toto+Qwen-3-4B	3.343×10^{-4}	8569.1	5.85	40.60	0.670	0.988	0.992	0.400	0.936
Toto+Qwen-3-4B+Thinking	7.101×10^{-4}	9120.4	6.42	44.10	0.632	0.978	0.965	0.388	0.894

The Scale Ablation

- Preserving absolute scale **matters** — a lot
- **Consistent gains across all 5 architectures and 6 tasks**
- **Autoformer:** +30 points on root cause, +41 points on traffic classification
- Absolute magnitude carries diagnostic information that temporal patterns alone can't recover

Model	Root Cause	Anom. Det. (F1)	Location	Traffic	Congestion	Mobility
TimesNet	0.688 → 0.756	0.462 → 0.740	0.913 → 0.923	0.993 → 0.997	0.989 → 0.991	0.972 → 0.992
Informer	0.656 → 0.800	0.451 → 0.852	0.915 → 0.929	0.993 → 0.996	0.992 → 0.986	0.959 → 0.994
FEDformer	0.492 → 0.620	0.330 → 0.713	0.873 → 0.926	0.975 → 0.988	0.776 → 0.845	0.972 → 0.972
Autoformer	**0.280 → 0.584**	**0.318 → 0.784**	**0.118 → 0.424**	**0.578 → 0.989**	**0.548 → 0.824**	**0.622 → 0.831**
NS Transformer	0.516 → 0.708	0.483 → 0.804	0.891 → 0.944	0.974 → 0.997	0.995 → 0.994	0.950 → 0.981

Takeaways

- TelecomTS opens up observability time series research
- First large-scale, de-anonymized, scale-preserving observability dataset
- Multi-modal Q&A with reasoning traces beyond forecasting
- Scale awareness is essential for foundation models in observability settings

https://huggingface.co/datasets/AliMaatouk/TelecomTS

https://github.com/Ali-maatouk/TelecomTS

Thank you!

Yale SCHOOL OF ENGINEERING
AND APPLIED SCIENCE