Can LLMs Understand Time Series Anomalies? *[ICLR 25]*

Zihao Zhou, Rose Yu
Dept of Computer Science and Engineering, University of California, San Diego
La Jolla, CA 92093, USA

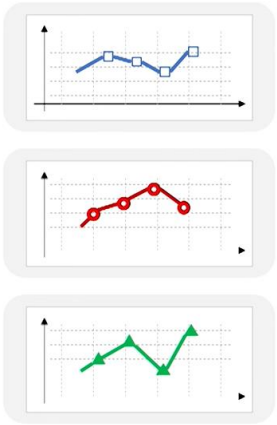
Sunghyun Ahn
skd@yonsei.ac.kr

<2025/03/10>

1 Time Series Anomaly Detection

Time Seies Anomaly Detection

- **시계열 데이터:** 시간 순서대로 나열된 테이블 데이터 (데이터 수(N) x 데이터 측정 시간(T), 변수의 개수에 따라 단변량, 다변량)
- **이상 탐지 활용 분야:** 이상 금융 탐지, 네트워크 상태 점검, 건강 상태 확인, 교통 원활 정도 확인 등
- 본 논문은 시계열 이상 탐지에서 LLM의 능력에 대한 포괄적인 분석을 제공함 *(단순한 단변량 시계열 데이터 활용)*



	00:00:00	00:01:00	00:02:00	00:03:00	...
A	102	100	90	130	...

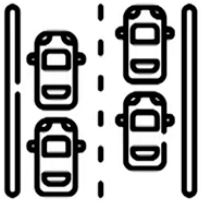
	00:00:00	00:01:00	00:02:00	00:03:00	...
B	0.1	1.5	2.3	0.3	...

	00:00:00	00:01:00	00:02:00	00:03:00	...
C	1002	1200	960	1300	...

주식

인터넷 속도

심장 박동



시간당 교통량

1 Time Series Anomaly Detection

Anomaly Pattern Classification

- **Out-of-range Anomalies:** 정상 값의 threshold를 크게 벗어나는 비정상 유형
- **Contextual Anomalies:** 특정 구간에서 맥락이 변경되는 비정상 유형 (Trend, Frequency, Point)

1. **Trend:** 갑작스럽게 acceleration, deceleration 또는 reversal이 발생하는 현상

2. **Frequency:** 주기적 패턴이 예상과 다르게 변화하는 현상
(특정 구간에서 패턴이 변형됨, 주기가 짧아지거나 길어짐)

3. **Point:** 전반적 패턴을 유지하지만 특정 데이터 포인트가 기대값에서 벗어나는 현상

4. **Out-of-Range:** 데이터 값이 normal range를 크게 벗어나는 현상

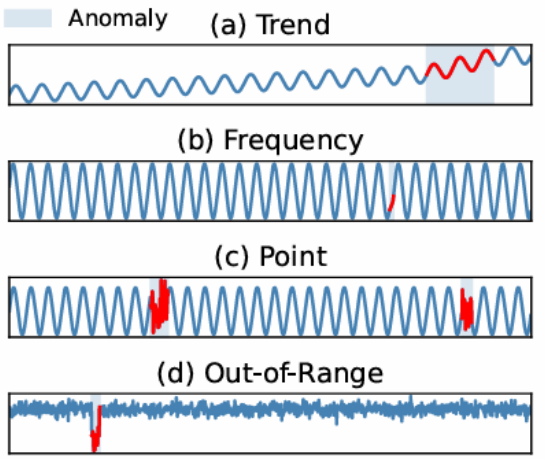


Figure 1: Example time series with different anomaly types, with anomalous regions highlighted in red.

1 Time Series Anomaly Detection

Zero-Shot and Few-Shot Anomaly Detection

- n개의 라벨링된 time-series를 입력받아, 새로운 time-series에 대한 비정상 라벨을 예측함 (time point or time interval)
- M-LLMs (e.g. Qwen, LLaMA, Gemini, GPT)를 활용하면, time-series 데이터를 text 혹은 image 형태로 입력하여 예측 가능함

$$\{y_1, y_2, \ldots, y_T\} = g(X_{new}, \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\})$$

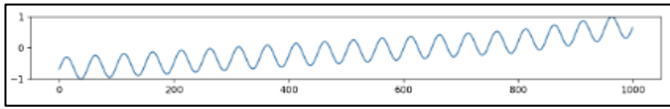
`'[{"start": 875, "end": 975}]\n'`

Y_{new}

Detect ranges of anomalies in this time series, in terms of the x-axis coordinate.
List one by one, in JSON format.
If there are no anomalies, answer with an empty list [].

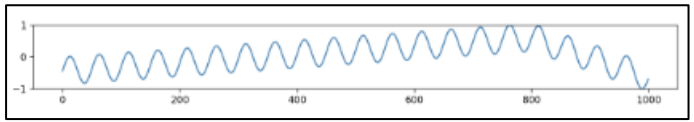
Output template:
[{"start": ..., "end": ...}, {"start": ..., "end": ...}...]

Prompt



[{"start": 800, "end": 975}]

(X_1, Y_1)



X_{new}

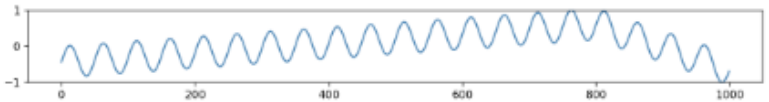
1 Time Series Anomaly Detection

Multimodal LLMs (M-LLMs)

- Text는 raw numerical values를 LLM에 입력함, Image는 matplotlib을 통해 제작된 visualized time series를 입력함
- 시계열 데이터를 이미지로 처리할 때 이상 탐지에서 훨씬 더 뛰어난 성능을 보인다는 것을 발견함

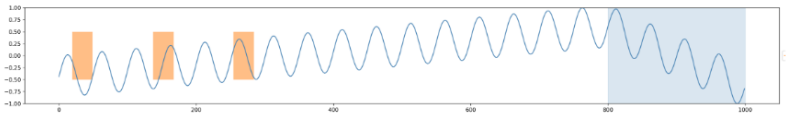
USER:

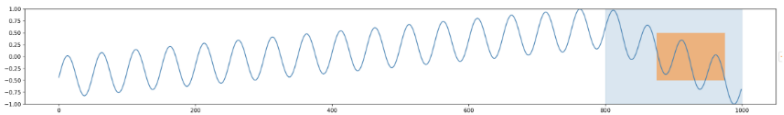
-0.44 -0.25 -0.1 -0.01 0.02 -0.04 -0.15 -0.32 -0.49 -0.66 -0.77 -0.82 -0.8 -0.7 -0.55 -0.36 -0.18 -0.03 0.06 0.08 0.03 -0.09 -0.26 -0.43 -0.59
-0.71 -0.76 -0.73 -0.63 -0.47 -0.29 -0.11 0.04 0.13 0.15 0.09 -0.03 -0.2 -0.38 -0.53 -0.65 -0.69 -0.66 -0.56 -0.4 -0.22 -0.04 0.11 0.2 0.21 0.15
0.03 -0.14 -0.32 -0.47 -0.58 -0.63 -0.59 -0.49 -0.33 -0.15 0.03 0.18 0.26 0.27 0.21 0.09 -0.08 -0.26 -0.41 -0.52 -0.56 -0.52 -0.42 -0.26 -0.07
0.1 0.25 0.33 0.34 0.27 0.15 -0.02 -0.2 -0.35 -0.46 -0.49 -0.46 -0.35 -0.19 -0.0 0.18 0.32 0.4 0.4 0.34 0.21 0.04 -0.14 -0.29 -0.39 -0.43 -0.39
-0.28 -0.12 0.07 0.25 0.39 0.46 0.47 0.4 0.27 0.1 -0.08 -0.23 -0.33 -0.36 -0.32 -0.21 -0.04 0.14 0.32 0.46 0.53 0.53 0.46 0.33 0.16 -0.02
-0.17 -0.27 -0.3 -0.25 -0.14 0.03 0.21 0.39 0.52 0.6 0.6 0.52 0.39 0.22 0.04 -0.11 -0.21 -0.23 -0.18 -0.07 0.1 0.29 0.46 0.59 0.66 0.66 0.58
0.45 0.28 0.1 -0.05 -0.14 -0.17 -0.11 0.0 0.17 0.36 0.53 0.66 0.73 0.72 0.65 0.51 0.34 0.16 0.01 -0.08 -0.1 -0.05 0.08 0.24 0.43 0.6 0.73 0.8
0.79 0.71 0.57 0.39 0.22 0.08 -0.02 -0.03 0.02 0.15 0.31 0.5 0.67 0.8 0.86 0.85 0.77 0.63 0.45 0.28 0.14 0.05 0.03 0.09 0.22 0.39 0.57 0.74
0.87 0.93 0.92 0.83 0.69 0.51 0.34 0.2 0.11 0.1 0.16 0.29 0.46 0.64 0.81 0.94 1.0 0.98 0.89 0.75 0.57 0.4 0.26 0.18 0.17 0.23 0.36 0.53 0.7
0.84 0.94 0.97 0.93 0.81 0.64 0.44 0.24 0.08 -0.03 -0.06 -0.02 0.09 0.23 0.39 0.54 0.63 0.66 0.61 0.49 0.32 0.12 -0.08 -0.24 -0.34 -0.37
-0.33 -0.22 -0.07 0.09 0.23 0.32 0.35 0.3 0.17 0.0 -0.2 -0.4 -0.56 -0.66 -0.69 -0.64 -0.53 -0.38 -0.22 -0.08 0.01 0.03 -0.02 -0.14 -0.32 -0.52
-0.72 -0.87 -0.97 -1.0 -0.95 -0.84 -0.69




```
In [37]: response = send_openai_request(request, model='gemini-1.5-flash')
         response

2024-10-12 06:29:38.504 | DEBUG    | gemini_api:send_gemini_request:93 - API key: ****************
GOM

Out[37]: ```json\n[{"start": 19, "end": 49}, {"start": 137, "end": 167}, {"start": 254, "end": 284}]\n```
```

```
In [24]: response = send_openai_request(request, model='gemini-1.5-flash')
         response

2024-10-12 06:26:05.999 | DEBUG    | gemini_api:send_gemini_request:93 -
GOM

Out[24]: '[{"start": 875, "end": 975}]\n'
```





Text **Image**

1 Time Series Anomaly Detection

Hypotheses

LLM의 이상 탐지 능력을 이해하기 위해, 여러 가설을 설정하고, 해당 가설을 테스트함

Hypothesis 1 (Tan et al., 2024) on Chain-of-Thought (CoT) Reasoning

LLMs do not benefit from engaging in step-by-step reasoning about time series data.

1. LLM은 시계열 데이터를 단계별로 추론해도 성능 향상 x

Hypothesis 2 (Gruver et al., 2023) on Repetition Bias

LLMs' repetition bias (Holtzman et al., 2020) corresponds precisely to their ability to identify and extrapolate periodic structure in the time series.

2. LLM의 반복 편향 (=)
시계열 데이터에서 주기적 구조를 식별하는 능력

Hypothesis 3 (Gruver et al., 2023) on Arithmetic Reasoning

LLMs' ability to perform addition and multiplication (Yuan et al., 2023) maps onto extrapolating linear and exponential trends.

3. LLM의 덧셈, 곱셈 수행 능력
-> 시계열의 선형 및 지수적 경향을 추론

Hypothesis 4 (Dong et al., 2024) on Visual Reasoning

Time series anomalies can be more easily detected as visual input rather than text input.

4. 텍스트 입력 << 시각적 입력

Hypothesis 5 on Visual Perception Bias

LLMs exhibit similar detection limitations to human perceptual biases, e.g., in acceleration perception when analyzing visual time series representations.

5. 인간의 지각적 한계와 유사한 결과를 보임

Hypothesis 6 on Long Context Bias

LLMs perform better for time series with fewer tokens, even if there is information loss.

6. 토큰 수가 적은 시계열 데이터가 더 나은 성능을 보임

1 Time Series Anomaly Detection

Hypotheses (1, 2)

Retained Hypothesis 1 on CoT Reasoning

No evidence is found that explicit reasoning prompts via CoT improve LLMs' performance in time series analysis.

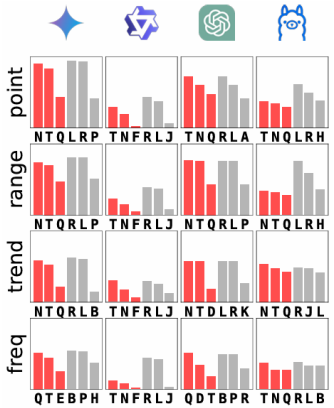


Figure 3: Reflexive (prompt that induces reasoning) / Reflective (prompt asks for direct answer),

CoT를 사용할 경우 성능이 오히려 감소함 -> 인간의 추론 과정과 LLM의 추론 과정은 다름

1) Perfect periodicity: $f(t+P) = f(t)$ for some period $P > 0$

2) Noisy periodicity: $f(t+P) = f(t) + \epsilon(t)$ where $\epsilon(t) \sim \mathcal{N}(0, \sigma^2)$

Rejected Hypothesis 2 on Repetition Bias

LLMs' repetition bias does not explain their ability to identify periodic structures.

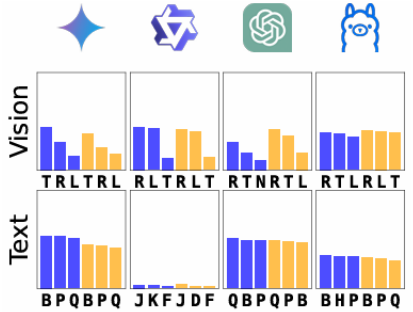


Figure 4: Clean (original time series) / Noisy (time series with minimal injected noise), Top 3 Affi-F1 variants per noise level

반복 편향은 LLM이 같은 패턴(e.g. 문장은 마침표로 끝남)을 반복한다는 의미임. 따라서 주기가 있는 시계열 데이터에서 같은 값이 반복될 때, LLM은 그 위치를 학습하여 반복적으로 예측할 수 있음.
이 때, 데이터에 noise를 추가하면 더 이상 반복적이지 않으므로, LLM의 반복 편향이 제대로 작동하지 않게 되며 예측에 어려움이 생길 것이라고 예상됨, 하지만 여전히 예측을 잘 함

1 Time Series Anomaly Detection

Hypotheses (3, 4)

Rejected Hypothesis 3 on Arithmetic Reasoning

The LLMs' understanding of time series is not related to its ability to perform arithmetic calculations.

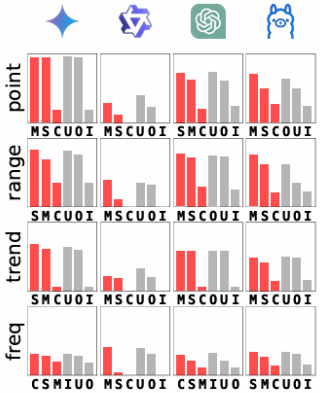


Figure 5: Calc (prompt with correct arithmetic example) / DysCalc (incorrect example), Top 3 Affi-F1 variants per mode

올바르지 않은 연산 결과를 in-context learning으로 미리 학습하여 LLM의 산술 능력을 강제로 떨어뜨림 -> 성능이 그래도 유지됨

Retained Hypothesis 4 on Visual Reasoning

Time series anomalies are better detected by M-LLMs as images than by LLMs as text.

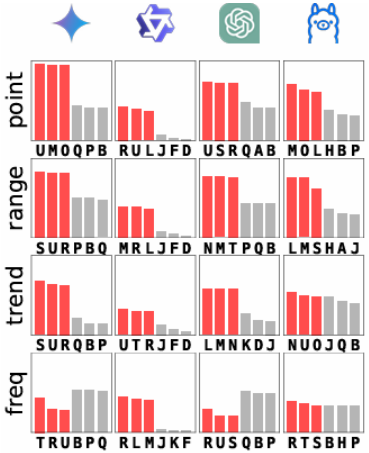


Figure 6: Vision (prompt with visualized time series) / Text (raw numerical prompt), Top 3 Affi-F1 variants per modality

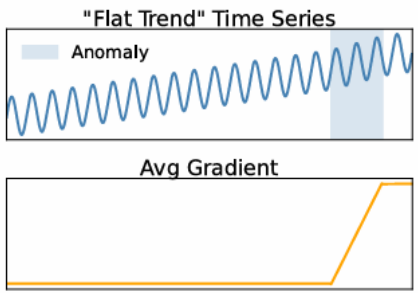
사람이 시계열 데이터를 텍스트보다 시각적으로 확인할 때 더욱 탐지를 잘하는 것처럼 LLM도 이미지에서 더욱 탐지를 잘 함

1 Time Series Anomaly Detection

Hypotheses (5, 6)

Rejected Hypothesis 5 on Visual Perception Bias

The LLM's understanding of anomalies is not consistent with human perception.



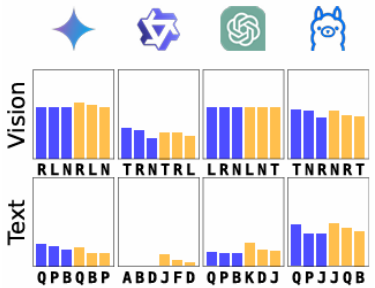


Figure 8: Flat Trend (see above for an example) / Trend (trend may reverse during anomalies), Top 3 Affi-F1 variants per dataset

"Flat Trend"는 사람이 판단하기 힘들지만 gradient에서 확연한 차이가 나는 데이터를 의미함. 시계열 데이터가 반전되는 구분하기 쉬운 "Trend" 데이터와 비교하여 LLM은 둘 다 높은 성능을 보임
-> 사람의 지각 능력과는 관련이 없음

Retained Hypothesis 6 on Long Context Bias

LLMs perform worse when the input time series have more tokens.

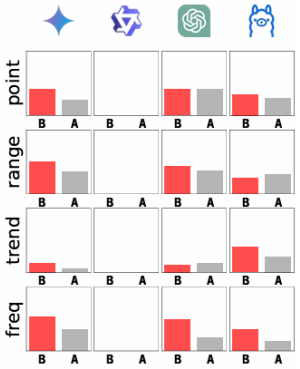


Figure 7: Subsampled (time series subsampled to be shorter) / Original, 0-shot raw text vs 30% text

1000개의 시간 토큰과 서브 샘플링된 300개의 시간 토큰의 성능을 비교한 결과, 오히려 300개에 대해서 더 높은 성능을 보임
-> LLM이 긴 시퀀스 처리에 어려움을 겪음

1 Time Series Anomaly Detection

Hypotheses (appendix)

Rejected Hypothesis 7 on Architecture Bias

LLMs' time series understanding vary significantly across different model architectures.

1. GPT-4o-mini
   - CoT 프롬프트 사용에 큰 차이 없음
   - frequency 이상 탐지에서 CoT 사용 시 성능 약간 향상됨

2. Qwen (7B)
   - 비전 프롬프트에서 좋은 성능을 보이지만, 텍스트 프롬프트에서 성능 저조함
   - CoT 사용 시 성능이 크게 하락됨

3. Gemini
   - GPT-4o와 유사한 성능을 보임
   - frequency 이상 탐지에서 비전 프롬프트를 이용해도 어려움

4. InternVL2-Llama3 (76B)
   - 비전 프롬프트와 텍스트 프롬프트의 성능 차이가 적음

1 Time Series Anomaly Detection

Conclusions

- LLM의 시계열 이상 탐지 능력을 종합적으로 조사함
- 전반적으로 인간과 처리 방식이 다르며, 인간처럼 시각 정보를 더욱 잘 활용한다는 특징을 발견함
- 시계열 이상 탐지 성능 향상을 위해 모델 선택 및 앙상블 방법이 중요함

→ LLM의 시계열 분석 능력을 조사하고, 이상 탐지 성능 향상을 위한 기법을 다룸

Thank You