BEDTime: A Unified Benchmark for Automatically Describing Time Series

Medhasweta Sen[1] Zachary Gottesman[1] Jiaxing Qiu[1] C. Bayan Bruss[2] Nam Nguyen[2] Tom Hartvigsen[1]
[1]University of Virginia [2]CapitalOne

https://huggingface.co/datasets/HartvigsenGroup/BEDTime
https://github.com/hartvigsen-group/bedtime

Forty-Third International Conference on Machine Learning

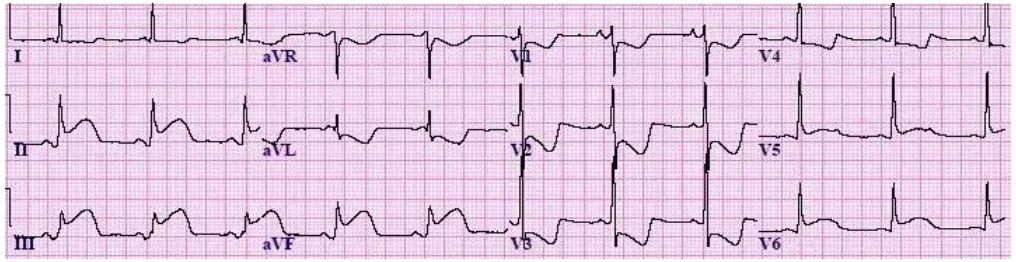
UNIVERSITY of VIRGINIA

CapitalOne

Multimodal time series data is everywhere

Multimodal time series data is everywhere

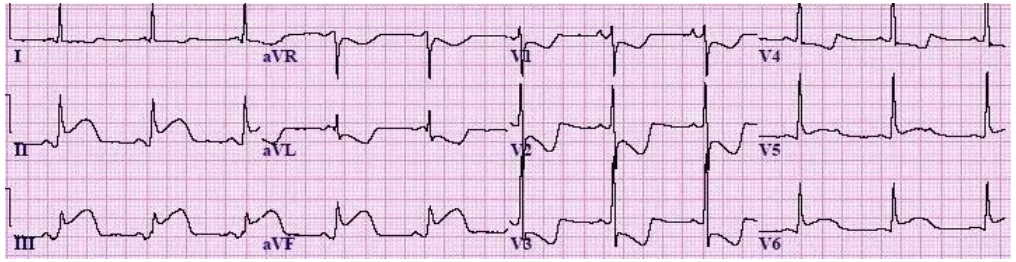
Healthcare



"58-year-old male with hypertension, high cholesterol, and smoking history. Presented with sudden chest pain, sweating, and shortness of breath."

Multimodal time series data is everywhere

Healthcare



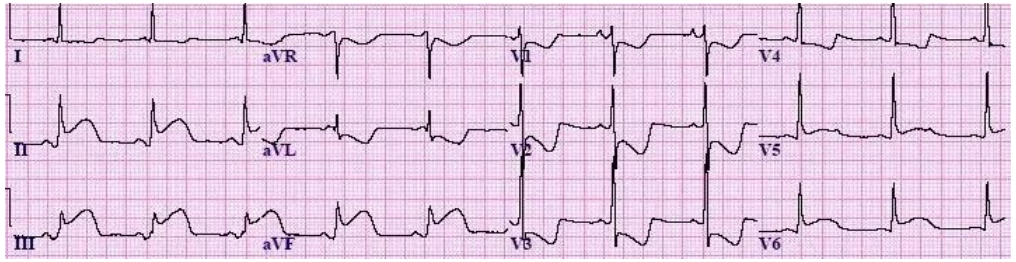
"58-year-old male with hypertension, high cholesterol, and smoking history. Presented with sudden chest pain, sweating, and shortness of breath."

Finance

"In 2022, Russia's invasion of Ukraine disrupted global energy and food supply chains. Oil and natural gas prices spiked, while shortages of wheat and fertilizer drove up food costs worldwide. These shocks pushed the U.S. CPI sharply higher, contributing to the highest inflation rates in four decades."

Multimodal time series data is everywhere

Healthcare

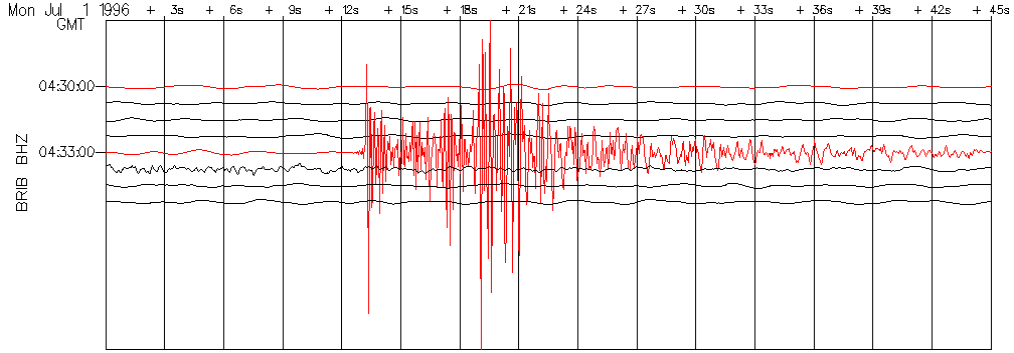


"58-year-old male with hypertension, high cholesterol, and smoking history. Presented with sudden chest pain, sweating, and shortness of breath."

Finance

"In 2022, Russia's invasion of Ukraine disrupted global energy and food supply chains. Oil and natural gas prices spiked, while shortages of wheat and fertilizer drove up food costs worldwide. These shocks pushed the U.S. CPI sharply higher, contributing to the highest inflation rates in four decades."

Climate Science



"Magnitude 6.7, Caused by a sudden rupture along the subduction zone where the Pacific Plate dives beneath the North American Plate. The megathrust displacement shifted the seafloor by several meters, unleashing a tsunami."

Rise of multi-modal time series foundational models

Rise of multi-modal time series foundational models

- **Large Language Models** read the series as text: `LLMTime` treats it as next-token prediction, `PromptCast` frames it as Q&A

Rise of multi-modal time series foundational models

- **Large Language Models** read the series as text: `LLMTime` treats it as next-token prediction, `PromptCast` frames it as Q&A
- **Time series and Language Models** feed the raw signal directly with a growing suite of models : `ChatTS`, `ChatTime`, `Time-LLM`, `SensorLM`, etc.

Rise of multi-modal time series foundational models

- **Large Language Models** read the series as text: `LLMTime` treats it as next-token prediction, `PromptCast` frames it as Q&A
- **Time series and Language Models** feed the raw signal directly with a growing suite of models : `ChatTS`, `ChatTime`, `Time-LLM`, `SensorLM`, etc.
- **Vision Language Models** aligns the plotted image of the time series with language.

Rise of multi-modal time series foundational models

- **Large Language Models** read the series as text: `LLMTime` treats it as next-token prediction, `PromptCast` frames it as Q&A
- **Time series and Language Models** feed the raw signal directly with a growing suite of models : `ChatTS`, `ChatTime`, `Time-LLM`, `SensorLM`, etc.
- **Vision Language Models** aligns the plotted image of the timeseries with language.
- Three modalities, many architectures, with claims about "reasoning" over time series

The Benchmark Gap

The Benchmark Gap

- **Reasoning suites:** TimeSeriesExam, Time-MQA

The Benchmark Gap

- **Reasoning suites:** TimeSeriesExam, Time-MQA
- **Domain QA:** ITFormer (aero-engines), MTBench (finance + weather)

The Benchmark Gap

- **Reasoning suites:** TimeSeriesExam, Time-MQA
- **Domain QA:** ITFormer (aero-engines), MTBench (finance + weather)
- **Retrieval and understanding:** TADACap, CaTS-Bench, QuAnTS

The Benchmark Gap

- **Reasoning suites:** TimeSeriesExam, Time-MQA
- **Domain QA:** ITFormer (aero-engines), MTBench (finance + weather)
- **Retrieval and understanding:** TADACap, CaTS-Bench, QuAnTS
- *But* — all target complex, downstream reasoning, mostly MCQ, often domain-locked, often with machine-generated text

The Benchmark Gap

- **Reasoning suites:** TimeSeriesExam, Time-MQA
- **Domain QA:** ITFormer (aero-engines), MTBench (finance + weather)
- **Retrieval and understanding:** TADACap, CaTS-Bench, QuAnTS
- *But* — all target complex, downstream reasoning, mostly MCQ, often domain-locked, often with machine-generated text

None test a fundamental capability: *Can a language model look at a time series and describe its salient structural features?*

What are Structural Features?

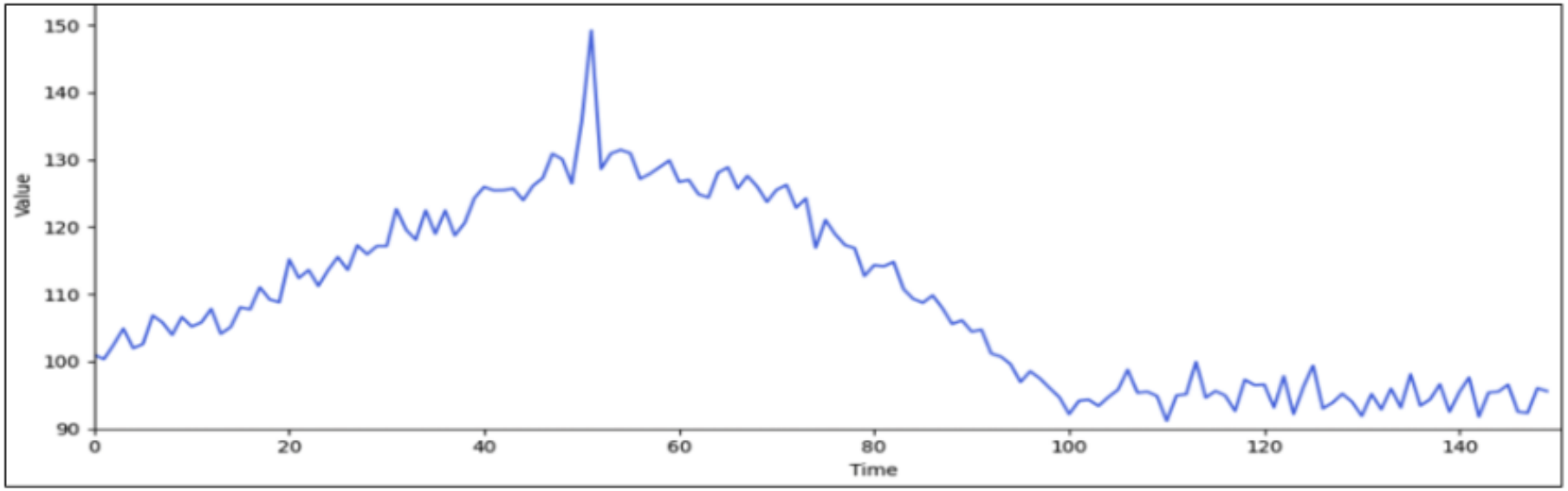
1. **Trend**
   - Directional movements: Up, Down
2. **Seasonality & Cyclical Patterns**
   - Fixed-period (constant or varying amplitude), Shifting period, Multiple seasonality
3. **Regime Switching / Structural Breaks**
   - Regime changes, Parameter shifts
4. **Random Processes**
   - White Noise, Random Walk
5. **Volatility**
   - Constant, Increasing, Clustered, Leverage effect
6. **Anomalies**
   - Spikes, Step-spikes, Level shifts, Temporal disruptions

The Foundational Task BEDTime Evaluates

Raw Time Series

[100.99, 100.34, 102.52, 104.88, 101.98, 102.59, 106.83, 105.82, 103.96, 106.6, 105.2, 105.8, 107.83, 104.13, 105.12, 108.06, 107.77, 111.04, 109.2, 108.81, 115.18, 112.41, 113.6,..., 92.66, 97.29, 96.5, 96.58, 93.18, 97.81, 92.2, 96.17, 99.38, 93.02, 93.87, 95.2, 93.99, 91.9, 95.14, 92.88, 95.95, 93.16, 98.1, 93.43, 94.36, 96.63, 92.54, 95.45, 97.61, 91.79, 95.37, 95.52, 96.56, 92.53, 92.36, 96.04, 95.59]

Raw Time Series Plotted



Time Series Description

This time series shows a steady upward trend with moderate noise, peaking in a sharp anomaly near the midpoint, followed by a gradual decline and transition into a noisy stationary regime.

Introducing BEDTime

Introducing BEDTime

3 Tasks

Introducing BEDTime

3 Tasks

Recognition

Introducing BEDTime

3 Tasks

Recognition

Differentiation

Introducing BEDTime

3 Tasks

Recognition

Differentiation

Generation

Introducing BEDTime

3 Tasks

17 Models|3 Modalities

Recognition

Differentiation

Generation

Introducing BEDTime

3 Tasks

Recognition

Differentiation

Generation

17 Models|3 Modalities

9 Large Language Models

Introducing BEDTime

3 Tasks

Recognition

Differentiation

Generation

17 Models|3 Modalities

9 Large Language Models

6 Vision Language Models

Introducing BEDTime

3 Tasks	17 Models|3 Modalities
Recognition	9 Large Language Models
Differentiation	6 Vision Language Models
Generation	2 Time Series Language Models

Introducing BEDTime

3 Tasks

17 Models|3 Modalities

5 Datasets

Recognition

Differentiation

Generation

9 Large Language Models

6 Vision Language Models

2 Time Series Language Models

Introducing BEDTime

3 Tasks

Recognition

Differentiation

Generation

17 Models|3 Modalities

9 Large Language Models

6 Vision Language Models

2 Time Series Language Models

5 Datasets

46,843 unique time series-text description pairs

Introducing BEDTime

3 Tasks

Recognition

Differentiation

Generation

17 Models|3 Modalities

9 Large Language Models

6 Vision Language Models

2 Time Series Language Models

5 Datasets

46,843 unique time series-text description pairs (90.44% real-world)

Key Findings

Key Findings

- Vision-language models are the strongest modality on every task.

Key Findings

- Vision-language models are the strongest modality on every task.
- LLMs perform worst.

Key Findings

- Vision-language models are the strongest modality on every task.
- LLMs perform worst.
- Time series-language models are competitive with similarly-sized LLMs but still trail behind similarly-sized VLMs.

Key Findings

- Vision-language models are the strongest modality on every task.
- LLMs perform worst.
- Time series-language models are competitive with similarly-sized LLMs but still trail behind similarly-sized VLMs.
- Input representation matters more than scale.

Key Findings

- Vision-language models are the strongest modality on every task.
- LLMs perform worst.
- Time series-language models are competitive with similarly-sized LLMs but still trail behind similarly-sized VLMs.
- Input representation matters more than scale.
- Even the best models fall short on these fundamental tasks, especially on noisy real-world data, leaving real room for improvement.

Robustness Experiments Finding

Robustness Experiments Finding

- **Sequence length:** LLM accuracy drops as series get longer, TSLMs are more robust in comparison.

Robustness Experiments Finding

- **Sequence length:** LLM accuracy drops as series get longer, TSLMs are more robust in comparison.
- **Missing data:** LMs are stable under < 25% missingness but collapses once > 50% of the time series are missing.

Robustness Experiments Finding

- **Sequence length:** LLM accuracy drops as series get longer, TSLMs are more robust in comparison.
- **Missing data:** LMs are stable under < 25% missingness but collapses once > 50% of the time series are missing.
- **Amplitude scaling:** All models are overall robust.

Robustness Experiments Finding

- **Sequence length:** LLM accuracy drops as series get longer, TSLMs are more robust in comparison.
- **Missing data:** LMs are stable under < 25% missingness but collapses once > 50% of the time series are missing.
- **Amplitude scaling:** All models are overall robust.
- **Additive noise:** All models degrades, with LLMs showing the fastest decline followed by VLMs and TSLMs

Robustness Experiments Finding

- **Sequence length:** LLM accuracy drops as series get longer, TSLMs are more robust in comparison.
- **Missing data:** LMs are stable under < 25% missingness but collapses once > 50% of the time series are missing.
- **Amplitude scaling:** All models are overall robust.
- **Additive noise:** All models degrades, with LLMs showing the fastest decline followed by VLMs and TSLMs
- **Image quality (VLMs):** Performance falls as plots get blurrier; frontier VLMs show slower decline.

Robustness Experiments Finding

- **Sequence length:** LLM accuracy drops as series get longer, TSLMs are more robust in comparison.
- **Missing data:** LMs are stable under < 25% missingness but collapses once > 50% of the time series are missing.
- **Amplitude scaling:** All models are overall robust.
- **Additive noise:** All models degrades, with LLMs showing the fastest decline followed by VLMs and TSLMs
- **Image quality (VLMs):** Performance falls as plots get blurrier; frontier VLMs show slower decline.
- **Chain-of-thought:** Consistent but modest gains for LLMs.

Conclusion

Conclusion

- **First unified, fully open benchmark** for the *foundational* task of describing time series with 46,843 pairs, **90%+ real-world**, 3 tasks × 3 modalities × 5 datasets. Code + data public.

Conclusion

- **First unified, fully open benchmark** for the *foundational* task of describing time series with 46,843 pairs, **90%+ real-world**, 3 tasks × 3 modalities × 5 datasets. Code + data public.
- **Modality is the deciding factor:** vision-language models lead on *every* task, *representation beats scale*, and the models built specifically for this still trail.

Conclusion

- **First unified, fully open benchmark** for the *foundational* task of describing time series with 46,843 pairs, **90%+ real-world**, 3 tasks × 3 modalities × 5 datasets. Code + data public.
- **Modality is the deciding factor:** vision-language models lead on *every* task, *representation beats scale*, and the models built specifically for this still trail.
- **A roadmap for future work:** even the best models break under realistic data and perturbations — exposing concrete failure modes the field can now target.