MTBench: A Multimodal Time Series Benchmark for Temporal Reasoning and Question Answering

(Jialin Chen *et al.*, 2024)

Presenter: Bingyue Li

Heidelberg University

Seminar: Cross-Temporal NLP

Outline

1. Motivation
2. Dataset Collection & Preprocessing
3. Task Curation
4. Experiments
5. Conclusion & Limitation

Motivation

- Real-world time-series tasks involve both:
  - numerical trends (e.g., stock prices, temperature)
  - textual narratives (e.g., news reports, weather summaries)
- Limitations of existing benchmarks:
  - Prediction-centric: under-evaluate reasoning-driven tasks
  - Weak cross-modal alignment: limited attention to semantic alignment
    $\rightarrow$ cannot handle cases where text contradicts the time series
  - Rigid dataset design

Introduction of MTBench

- **MTBench**: a multimodal time-series benchmark designed for
  - reasoning over time-series and text
  - evaluating cross-modal interactions
- Domains:
  - Finance
  - Weather
- Aligns time-series data with relevant text
- Supports diverse reasoning-intensive tasks beyond simple prediction

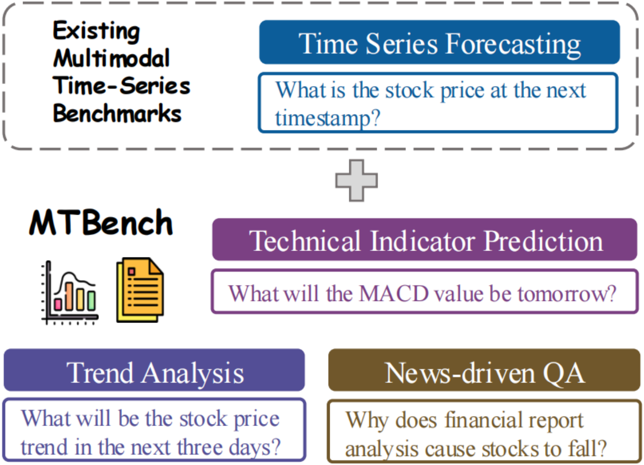


Figure 1: An overview of tasks in MTBench.

Data Collection and Preprocessing

Finance Data Collections

- Textual data
  - Extracted content, titles, stock names, and publication dates from financial news URLs
  - Curated a balanced 20K-news subset with diverse article lengths
  - Annotated with content type, temporal effect range, and sentiment
- Time-series data
  - Retrieved historical stock price time series using stock names
  - Sampled at different temporal granularities
  - Filtered low-quality samples

Weather Data Collections

- Time-series data
  - Collected temperature time-series from 50 U.S. airport stations using the GHCN-H dataset
  - Hourly data spanning 2003–2020
  - Airports chosen for higher data reliability and accuracy
- Textual data
  - Collected from the U.S. Storm Events Database
  - Covers severe weather conditions (e.g., hurricanes, floods, tornadoes)
  - Event records with event IDs and episode IDs
    - Event ID identifies a single occurrence
    - Episode ID groups related events under a larger weather system

Finance: News–Stock Alignment & Preprocessing

- Each news article is aligned with stock time series by
  - publication timestamp
  - corresponding stock name
- News sentiment is compared with the ground-truth future stock trend
- Dataset is split into two subsets:
  - **Consistent news**: sentiment aligns with future trend (80%)
  - **Misaligned news**: sentiment diverges from price movement (20%)
- Misaligned data serves as a crucial test of robustness to misleading text

Finance Data Overview

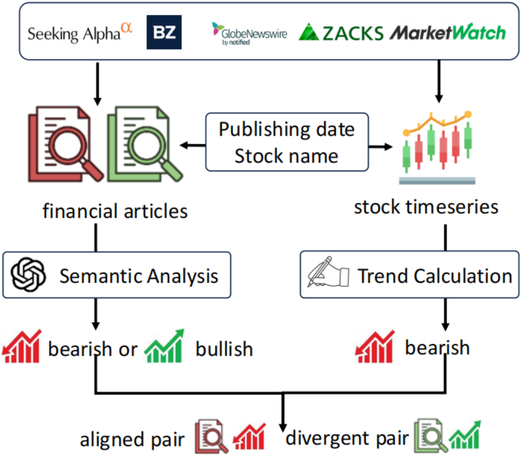


Figure 2: Pipeline of collecting news and time-series pairs.

Weather: Event–Record Alignment & Preprocessing

- Storm events are associated with nearby weather stations
  - nearest airport within a 50 km radius
  - events within the same episode IDs are merged into a record
  - episode-level textual reports are synthesized using LLMs
- Weather time series are anchored to the **episode end time** and standardized
  - fixed historical window
  - regular hourly resolution

Task Curation

Time-Series Forecasting Task

- **Task Setting**
  - Predict future time-series values from historical observations
- **Evaluation Metrics**
  - Finance: MAE, MAPE
  - Weather: MSE, MAE

Trend Analysis Task

- **Task Setting**
  - Classify the trend of a time series into discrete categories
  - Trend Calculation
    - Finance: percentage change between output window endpoints
    - Weather:
      - Past trend: slope of daily mean temperatures over input days
      - Future trend: difference between last input day and future day
  - Trend Label Binning
    - Finance: 3-way and 5-way classification
    - Weather: 3-way classification
- **Evaluation Metrics**
  - label classification accuracy

Table 1: Trend Label Binning for Financial and Weather Data

3-way Label	5-way Label	Financial Price	Weather Temperature
Negative	Bearish	$< -4\%$	Past: < -0.25 Future: < -0.5
	Warning	$-4\% \sim -2\%$	
Neutral	Neutral	$-2\% \sim 2\%$	Past: $-0.25 \sim 0.25$ Future: $-0.5 \sim 0.5$
Positive	Growth-Oriented	$2\% \sim 4\%$	Past: > 0.25 Future: > 0.5
	Bullish	$> 4\%$	

Technical Indicator Prediction

- **Task Setting**
  - Predict key indicators derived from output time series
  - Indicators capture higher-level properties beyond raw values
- **Indicators**
  - **Finance**:
    - Moving Average Convergence Divergence (MACD)
    - Upper Band of the Bollinger Bands
  - **Weather**:
    - Next-day max & min temperature
    - Next-day temperature difference
- **Evaluation Metrics**
  - MSE, MAE

News-driven Question Answering (QA)

- **Goal:** evaluate reasoning over **text + time series**
  - Requires understanding the news content and its relationship to future trends
- **Sub-task 1: Correlation Prediction**
  - Classify how a news article relates to future trends
  - Label schemes: 3-way (pos/neutral/neg) and 5-way (direction + magnitude)
  - Ground-truth labels generated by GPT-4o, based on actual price movements
- **Sub-task 2: Multiple-choice QA**
  - Select the correct statement grounded in news evidence and/or time-series facts
  - Incorrect options reflect plausible but flawed interpretations or false causal claims

Experiments

Experimental Setting

- **Baseline Models**
  - GPT-4o
  - Claude-Sonnet-3.5
  - Gemini-2.0-Flash
  - LLaMA 3.1-8B
  - DeepSeek-Chat
  - OpenAI-o1 (for certain finance related tasks)
- **Evaluation Protocol**
  - Models evaluated across all curated tasks
  - Two input settings:
    - Time-series only
    - Time-series + text
  - Two forecasting regimes:
    - Short-term
    - Long-term

Time-Series Forecasting

Table 2: Stock price forecasting under TS-only and TS+Text setting given 7-day or 30-day input.

	7-Day				30-Day			
	MAE		MAPE		MAE		MAPE	
	TS	w/ Text	TS	w/ Text	TS	w/ Text	TS	w/ Text
GPT-4o	1.687	1.596	**0.685**	2.544	2.387	2.338	3.739	3.520
Gemini	1.675	1.628	3.434	3.513	2.587	2.432	3.568	3.268
Claude	1.358	1.422	1.923	2.098	2.126	2.065	3.020	2.847
DeepSeek	1.753	1.720	2.085	2.135	2.357	2.134	3.482	3.305
OpenAI-o1	**1.058**	**0.982**	1.585	**1.424**	**1.842**	**1.703**	**2.598**	**2.240**

Table 3: Temperature forecasting under TS-only and TS+Text setting given 7-day or 14-day input. Llama fails to generate responses of the expected length in long-term scenarios.

	7-Day				14-Day			
	MSE		MAE		MSE		MAE	
	TS	w/ Text	TS	w/ Text	TS	w/ Text	TS	w/ Text
GPT-4o	**21.67**	**17.55**	**3.45**	**3.11**	45.59	40.43	4.65	4.49
Gemini	25.75	24.31	3.82	3.67	56.10	29.47	4.53	4.03
Claude	30.34	22.48	4.11	3.50	**32.01**	**25.08**	**4.24**	**3.75**
Llama3.1	51.62	48.54	5.66	5.26	/	/	/	/
DeepSeek	31.02	29.38	4.15	4.04	61.8	101.28	5.36	6.61

Semantic Trend Prediction

Table 4: Stock trend prediction accuracy with 3 and 5 trend labels on the news-stock pair dataset

	7-Day				30-Day			
	3-way		5-way		3-way		5-way	
	TS	*w/* Text	TS	*w/* Text	TS	*w/* Text	TS	*w/* Text
GPT-4o	40.93	42.81	34.18	36.45	34.90	47.35	19.85	30.58
Gemini	41.30	47.30	34.00	41.50	37.05	44.90	21.15	29.70
Claude	41.20	44.90	34.40	33.40	36.20	52.05	21.10	31.70
DeepSeek	40.53	45.12	32.85	35.60	35.50	48.26	20.70	29.55
OpenAI-o1	**47.50**	**60.99**	**37.50**	**54.41**	**39.25**	**59.12**	**22.50**	**43.24**

Table 5: Temperature trend prediction accuracy

	Past		Future	
	TS	*w/* Text	TS	*w/* Text
GPT-4o	69.47	**66.36**	23.07	43.54
Gemini	53.19	56.96	17.91	51.76
Claude	**70.44**	59.78	**33.23**	**56.87**
DeepSeek	22.61	26.49	16.89	25.17

Technical Indicator Prediction

Table 6: Stock indicator (MACD, and upper Bollinger Band (BB) prediction MSE under TS-only and TS+Text setting

	7-Day				30-Day			
	MACD		BB		MACD		BB	
	TS	w/ Text	TS	w/ Text	TS	w/ Text	TS	w/ Text
GPT-4o	0.430	0.365	1.450	1.082	1.003	0.897	2.521	2.068
Gemini	0.482	0.384	1.280	1.153	1.132	0.975	2.565	2.248
Claude	**0.241**	0.373	2.105	1.246	0.970	1.171	2.605	2.345
DeepSeek	0.435	0.352	1.526	1.187	1.053	1.072	2.486	2.201
OpenAI-o1	0.384	**0.246**	**1.025**	**0.687**	**0.823**	**0.586**	**2.015**	**1.523**

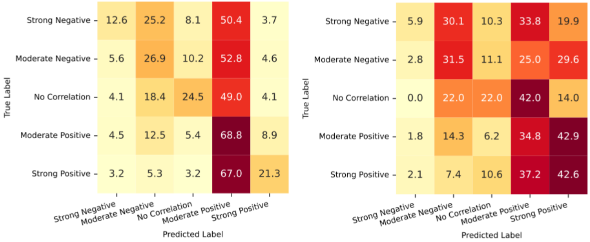
Table 7: Min/Max/Difference of Temperature Prediction

	Maximum				Minimum				Difference			
	MSE		MAE		MSE		MAE		MSE		MAE	
	TS	w/ Text	TS	w/ Text	TS	w/ Text	TS	w/ Text	TS	w/ Text	TS	w/ Text
GPT-4o	26.03	19.58	3.76	3.02	15.58	15.39	**2.89**	2.76	27.06	**18.84**	3.86	**3.20**
Gemini	25.98	**16.39**	3.77	**2.96**	16.20	16.27	2.94	2.93	35.72	23.21	4.40	3.63
Claude	**23.18**	18.69	**3.59**	3.21	**14.57**	**13.42**	2.73	**2.63**	**21.03**	19.10	**3.41**	3.26
Llama3.1	37.56	33.87	4.67	4.42	21.21	18.80	3.44	3.22	65.77	54.28	6.54	5.85
DeepSeek	33.90	32.82	4.45	4.38	18.39	17.25	3.16	3.05	49.28	44.99	5.51	5.24

News-driven Question Answering

Table 8: Accuracy Comparison on Different Tasks

	News-stock Correlation				News-driven MCQA			
	7-Day		30-Day		7-Day		30-Day	
	3-way	5-way	3-way	5-way	Finance	Weather	Finance	Weather
Llama3.1	17.2	9.4	32.1	16.4	35.4	33.8	61.8	29.1
Gemini	51.8	26.4	**59.6**	34.8	63.6	43.4	50.3	54.0
Claude	50.4	29.0	57.9	34.3	75.6	**51.8**	61.1	51.2
GPT-4o	**53.6**	**31.0**	57.6	34.6	65.1	41.7	52.8	44.8
DeepSeek	50.0	27.1	57.5	**35.0**	**77.6**	46.7	**69.3**	**57.3**



(a) GPT-4o (b) Gemini

Figure 8: Confusion map of correlation prediction results generated by (a) GPT-4o and (b) Gemini.

Conclusion & Limitation

Conclusion

- **Finance domain**
  - Short-term stock movements are noisy and difficult to correlate with news
  - Financial news is more informative for understanding long-term market trends
- **Weather domain**
  - Text is more helpful for future prediction
  - Less useful for retrospective analysis
- **Model behavior**
  - Errors are systematic rather than random, reflecting conservative reasoning
- **Overall insight**
  - Causal reasoning between text and time-series remains challenging for current LLMs

Limitations

- Limited cross-domain generalizability
  - Data alignment and preprocessing are highly domain-specific
  - Extending MTBench to new domains requires redesign of alignment rules
- Reliance on LLM-assisted annotation
  - Some labels and textual reports are generated or refined using LLMs
  - Potential annotation bias may affect evaluation robustness

Future direction

- Extend MTBench to additional domains (e.g., healthcare, social sciences)
- Fine-tuning strategies
- Architectural enhancements

Thank you!