FinSTaR: Towards Financial Reasoning with Time Series Reasoning Models

Anonymous Author(s)

Abstract

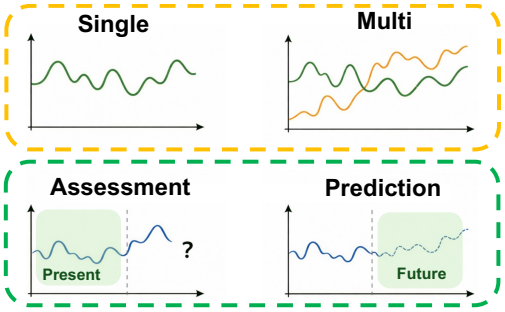
Time series (TS) reasoning models (TSRMs) have shown promising capabilities in general domains, yet they consistently fail on financial domain, which exhibit unique characteristics. We propose a general 2×2 capability taxonomy for TSRMs by crossing 1) *single-entity vs. multi-entity* analysis with 2) *assessment* of the current state vs. *prediction* of future behavior. We instantiate this taxonomy in the financial domain—where the distinction between deterministic assessment and stochastic prediction is particularly critical—as ten financial reasoning tasks, forming the **FinTSR-Bench** benchmark based on S&P stocks. To this end, we propose **FinSTaR** (**Fin**ancial Time **S**eries **T**hinking **a**nd **R**easoning), trained on FinTSR-Bench with distinct chain-of-thought (CoT) strategies tailored to each category. For *assessment*, which is *deterministic* (i.e., computable from observable data), we employ *Compute-in-CoT*, a programmatic CoT that enables models to derive answers directly from raw prices. For *prediction*, which is inherently *stochastic* (i.e., subject to unobservable factors), we adopt *Scenario-Aware CoT*, which generates diverse scenarios before making a judgment, mirroring how financial analysts reason under uncertainty. The proposed method achieves 78.9% average accuracy on FinTSR-Bench, substantially outperforming LLM and TSRM baselines. Furthermore, we show that the four capability categories are complementary and mutually reinforcing through joint training, and that Scenario-Aware CoT consistently improves prediction accuracy over standard CoT. Code is available at https://anonymous.4open.science/r/financial_tsrm-EE30.

1 Introduction

Time series (TS) reasoning models (TSRMs) have emerged as a promising direction for enabling LLMs to understand and reason over temporal data [10, 20, 35]. Unlike conventional forecasting, TSRMs produce interpretable reasoning chains that explain *why* a particular conclusion follows from the data. However, existing TSRMs are primarily developed on general domains and perform poorly on financial TS [18]. We find that existing models achieve accuracy barely above chance level on financial reasoning tasks, confirming that *existing TSRMs lack financial reasoning capabilities.*

Why do TSRMs fail on finance? Financial TS exhibit characteristics that fundamentally differ from general domains, including volatility clustering, momentum and mean-reversion dynamics, support and resistance levels, and complex cross-asset dependencies. More importantly, financial reasoning involves two distinct modes that existing TSRMs fail to distinguish:

- **[1] Assessment** (*"What is the current state?"*): *Deterministic* reasoning that can be resolved by computing observable quantities from price data (e.g., measuring drawdown severity).
- **[2] Prediction** (*"What will happen in the future?"*): *Probabilistic* reasoning under uncertainty, where outcomes depend on unobservable external factors (e.g., earnings announcements or macroeconomic shifts), and even a perfect model cannot achieve 100% accuracy.



[1] Can TSRM analyze single / multi stock(s)?

[2] Can TSRM assess / predict the values of stock?

Figure 1: Capability taxonomy for TSRMs.

Table 1: Overview of FinTSR-Bench. The proposed ten financial reasoning tasks are organized along two axes: 1) *Data* scope (single-stock vs. multi-stock) and 2) *Temporal* scope (assessment vs. prediction).

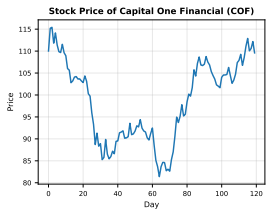
	Single	Multi
Assessment	Drawdown, Volatility, Trend	Correlation
Prediction	Event Response, Support/Resistance, Drawdown Recovery, Volatility Forecast	Relative Performance, Pair Convergence

However, no existing TSRM has been specifically designed for the financial domain or distinguishes between these two reasoning modes. We propose a general 2×2 capability taxonomy for TSRMs along two orthogonal axes: (data scope) *single-entity vs. multi-entity* and (temporal scope) *assessment vs. prediction* (Figure 1). While this taxonomy is applicable to TSRMs in any domain, we instantiate it in the financial domain, where the distinction between deterministic assessment and stochastic prediction is particularly critical, as ten financial reasoning tasks (Table 1). To support training and evaluation, we construct **FinTSR-Bench** from 250 S&P 500 stocks spanning 2010–2025, with ~35K training samples and three out-of-distribution (OOD) test splits for generalization evaluation.

To this end, we propose **FinSTaR** (**Fin**ancial Time **S**eries **T**hinking **a**nd **R**easoning), the first TSRM designed for the financial domain, as shown in Table 2. Trained on FinTSR-Bench, FinSTaR employs different chain-of-thought (CoT) strategies for assessment and prediction. Specifically, since assessment is deterministic and prediction is inherently stochastic, we use **Compute-in-CoT** for assessment tasks to derive answers from observable prices. For prediction,

Table 2: Comparison with related works. We compare across 1) domain, 2) input modality, 3) task, 4) CoT, and 5) uncertainty modeling. FinSTaR is the first TSRM designed for financial domain.

Work	Venue	Contribution		[1] Domain	[2] Input		[3] Task	[4] CoT	[5] Uncertainty
		Benchmark	Model		TS	Text			
S2TS-LLM [5]	NeurIPS'25		✓	General	✓		Analysis		
Thoth [35]	arXiv'26		✓		✓		Understanding		
PATRA [20]	arXiv'26		✓		✓		QA	✓	
SenTSR-Bench [3]	AISTATS'26	✓	✓		✓		Reasoning	✓	
TimeOmni-1 [10]	ICLR'26	✓	✓		✓	✓	Reasoning QA	✓	
FinTSB [16]	arXiv'25	✓		Finance	✓		Forecasting		
FinBen [30]	NeurIPS'24	✓				✓	NLP tasks		
CausalStock [24]	NeurIPS'24		✓		✓	✓	Forecasting		
FinZero [27]	arXiv'25	✓	✓			✓	Forecasting		✓
FinSTaR (Ours)		✓	✓		✓	✓	**Reasoning QA**	✓	✓



(Single-Assessment) Volatility Regime: *Classify recent vs. overall volatility ratio. GT: (C) High.*

Qwen2.5-7B (A) ✗

```
Calculate the overall volatility [...] Calculate the recent volatility [...] Compute the ratio [...] Let's perform these calculations.
```

▹ *No computation; directly outputs answer.*

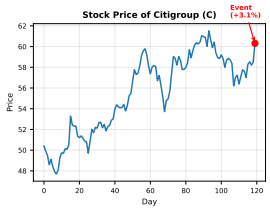
TimeOmni-1 (B) ✗

```
Estimate recent SD ≈ 2–3. Estimate overall SD ≈ 2–3. Ratio ≈ 1.0 ⇒ Normal.
```

▹ *Rough estimates; actual recent SD is 10.75%.*

FinSTaR (Ours) (C) ✓

```
Step 1: Compute daily returns.
Step 2: Overall volatility = 4.51%.
Step 3: Recent 20-day vol = 10.75%.
Step 4: Ratio = 10.75/4.51 = 2.38.
Recent vol well above long-term avg.
⇒ (C) High volatility.
```



(Single-Prediction) Event Response: *Predict mean-reversion or persistence after a shock. GT: (B) Persistence.*

Qwen2.5-7B (A) ✗

```
The overall pattern suggests the stock tends to revert to a mean after sharp moves. [...] the most likely outcome is that the price will move back toward pre-event levels.
```

TimeOmni-1 (A) ✗

```
The last value is 148.40, a significant jump from the previous day's low. This is an outlier [...] mean reversion is the most plausible scenario.
```

FinSTaR (Ours) (B) ✓

```
Step 1: Event: Positive shock, z=4.52, +13.70%.
Step 2: Pre-event trend: downward.
Step 3: Scenario analysis:
- Base: Shock reflects genuine new info.
- Adverse: Mean-reversion forces kick in.
- Favorable: Additional news amplifies move.
Step 4: 2/3 support Persistence.
Step 5: Judgment ⇒ (B).
```

Figure 2: Qualitative comparison. Two simplified examples where only FinSTaR answers correctly. Baselines either fail to compute precise quantities (Volatility Regime) or default to heuristic reasoning (Event Response), while FinSTaR's CoT produces grounded, step-by-step reasoning. ***Gray italic comments (▹) are editorial annotations, not model outputs.***

Scenario-Aware CoT generates multiple scenarios before making a probabilistic judgment. As illustrated in Figure 2, existing models [10, 23] either fail to compute precise quantities or rely on heuristic reasoning, while FinSTaR's structured CoT produces correct answers through grounded, step-by-step reasoning. Our main contributions are summarized as follows:

- We propose a 2×2 **capability taxonomy** for TSRMs by crossing *assessment* vs. *prediction* with *single-* vs. *multi-entity* analysis, and instantiate it as ten financial reasoning tasks.
- We construct **FinTSR-Bench** from 250 S&P 500 stocks (2010–2025) with 35K training samples and three OOD test sets for generalization evaluation, each containing 10K samples.
- We propose **FinSTaR**, trained with different CoT strategies tailored to each reasoning mode: Compute-in-CoT for assessment tasks and Scenario-Aware CoT for prediction tasks.
- FinSTaR achieves 78.9% overall accuracy, outperforming 15+ baselines, including LLMs and TSRMs. Through systematic ablation, we show that the four capability categories are complementary and mutually reinforcing under joint training.

2 Related Work

Time series reasoning models (TSRMs). Recent work extends LLMs beyond forecasting toward TS reasoning. TimeOmni-1 [10] introduces TSR-Suite with human-annotated CoT but focuses on non-financial domains. Thoth [35] proposes mid-training for TS understanding, PATRA [20] addresses pattern-aware alignment, and SenTSR-Bench [3] injects domain knowledge for TS reasoning. Merrill et al. [18] formalize TS reasoning as QA over numerical sequences, while S2TS-LLM [5] uses spectral symbolization to bridge TS and language representations. More recent work explores tool-augmented TSRMs and RL-based reasoning [14, 34], as well as cross-domain temporal reasoning benchmarks [28]. However, *all existing TSRMs target general domains* and none has been designed for financial TS, which require distinct reasoning about market-specific phenomena.

LLMs in financial domain. FinTSB [16] benchmarks forecasting on stocks but lacks reasoning tasks. FinBen [30] evaluates LLMs across financial NLP datasets but targets text-based tasks rather than raw price TS. News-to-Forecast [25] integrates event analysis

Table 3: Task specifications of FinTSR-Bench. Ten tasks span four capability categories across two axes: (data) single-stock vs. multi-stock and (temporal) assessment vs. prediction. Matching symbols (†, ‡, §) denote assessment–prediction pairs that share the same underlying financial skill.

Task category		Task	Classes	Description
Type	# Stock			
Assessment	Single	Drawdown†	4	Classify peak-to-trough decline severity
		Volatility Regime‡	3	Classify recent vs. overall volatility ratio
		Trend Direction	5	Classify 120-day cumulative return regime
	Multi	Correlation§	3	Classify return correlation of two stocks
Prediction	Single	Event Response	2	Mean-reversion or persistence after extreme move
		Support/Resistance	2	Breakout or bounce near key technical level
		Drawdown Recovery†	2	Recovery toward peak or further decline
		Volatility Forecast‡	2	Volatility increase or decrease
	Multi	Relative Performance	2	Which stock delivers higher returns
		Pair Convergence§	2	Price spread converges or diverges

with reflection and CausalStock [24] leverages LLMs as denoised news encoders, but both require textual inputs rather than reasoning from raw prices. No existing work combines *raw financial TS* with *reasoning QA* and *CoT supervision*, the gap that FinSTaR addresses. Recent work on financial verbal reasoning [17] and multimodal TS reasoning [6] explores complementary directions but does not address the assessment–prediction distinction central to our approach. A comparison with related work is shown in Table 2.

CoT for structured data. CoT [29] improves LLM reasoning on math [7] and science tasks, and STaR [31] bootstraps reasoning from self-generated rationales. Program-of-Thought (PoT) [4] executes code for arithmetic faithfulness, and verifiable-reward RL approaches [26] use outcome correctness as training signal. For TS reasoning, *existing CoT methods rely on human annotation [10] or GPT-generated rationales*, which are difficult to scale. Our Compute-in-CoT similarly ensures arithmetic correctness but within text-based reasoning chains without requiring a code execution environment, making it compatible with standard LLM inference. Scenario-Aware CoT further extends CoT to stochastic prediction by reasoning over multiple scenarios, a direction not explored by PoT or verifiable-reward methods which focus on deterministic correctness.

3 FinTSR-Bench

We instantiate the proposed 2×2 taxonomy in the financial domain as **FinTSR-Bench**, both a training dataset and evaluation benchmark for Financial TSRMs. As shown in Table 3, the proposed ten tasks cover all four capability categories in our taxonomy.

We use the top 250 S&P 500 stocks by market capitalization as of January 2024, with daily closing prices from 2010–2025 (~3,750 trading days). The top 200 stocks serve as in-domain (ID) training data, while the remaining 50 are held out for OOD evaluation. As shown in Table 4, we split along both the stock universe and time period axes, yielding three test sets that vary ID vs. OOD conditions. Each task has ~3,500 training samples and ~1,000 test samples per split, totaling ~35,000 for training and ~10,000 per test set. All label definitions (prediction horizons, thresholds, and class distributions balanced via undersampling) are provided in Appendix A, along with task examples in Appendix D.

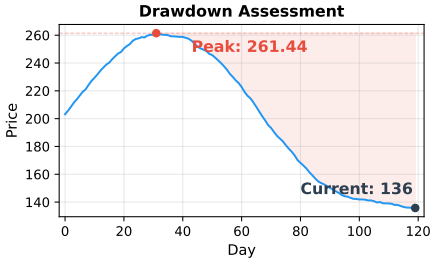
Table 4: Data splits of FinTSR-Bench. Three test sets systematically vary 1) stock universe and 2) time period to evaluate both ID and OOD.

Split	# Stocks	Period	Universe	Period
Train	200	2010–2022	–	–
Test A	200	2023–2025	ID	OOD
Test B	50	2010–2022	OOD	ID
Test C	50	2023–2025	OOD	OOD

4 Method: FinSTaR

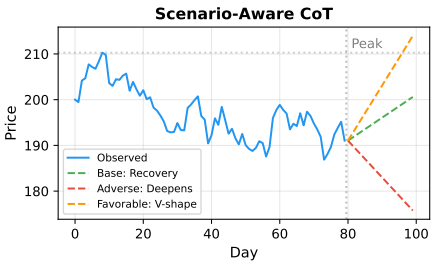
We train FinSTaR via supervised fine-tuning (SFT) on FinTSR-Bench data with structurally different chain-of-thought supervision for 1) assessment tasks and 2) prediction tasks.

Why two CoT strategies? Assessment and prediction tasks have fundamentally different epistemological properties. Assessment tasks (e.g., "What is the current drawdown?") have answers that are *fully determined* by the observable price data, and the model simply needs to compute the right quantity. Prediction tasks (e.g., "Will this drawdown recover?") have answers that depend on *future events* influenced by unobservable factors such as earnings surprises. Even with a perfect model, prediction accuracy cannot reach 100% because the necessary information is not contained in the price history alone. This fundamental distinction between *deterministic* and *probabilistic* reasoning motivates our design of two complementary CoT strategies tailored to each reasoning mode: Compute-in-CoT for *deterministic* assessment and Scenario-Aware CoT for *stochastic* prediction.




```
Step 1: Peak = 261.44 (day 30).
Step 2: Current price = 136.00.
Step 3: Drawdown = (261.44−136.00)/261.44 = 48.0%
        48.0% > 20% ⇒ (D) Severe Decline (> 20%).
```

(a) Simplified example of Compute-in-CoT: Assessment task - Drawdown.




```
Step 1: Peak = 210.15 (day 8), Current = 191.08.
Step 2: Drawdown = 9.1%, momentum = −2.1%.
Step 3: Scenario analysis:
  • Base: Stabilizes, value buyers → recovery.
  • Adverse: Earnings revision → decline.
  • Favorable: Upgrade → V-shaped recovery.
Step 4: Consider all scenarios ⇒ recovery.
Step 5: Judgment: Recovery ⇒ (A).
```

(b) Simplified example of Scenario-Aware CoT: Prediction task - Drawdown Recovery.

Figure 3: Two CoT strategies (simplified). Assessment tasks use Compute-in-CoT that extracts and computes quantities from observable prices. Prediction tasks use Scenario-Aware CoT that generates base/adverse/favorable scenarios before making a judgment. Full examples are in Appendix D.

4.1 Compute-in-CoT (Assessment Tasks)

For assessment tasks, the answer is deterministically computable from observable prices. We generate CoT chains programmatically by executing the same computation the model should learn. Each chain follows a three-phase structure (Figure 3a):

Phase 1. **Extract**: Identify relevant quantities from input prices (e.g., peak price, recent returns).

Phase 2. **Compute**: Perform the calculation (e.g., drawdown percentage, volatility ratio).

Phase 3. **Classify**: Map the computed result to the answer using the thresholds in the question.

This guarantees 100% correctness and unlimited scalability, as models never reference unobservable information at inference time.

4.2 Scenario-Aware CoT (Prediction Tasks)

For prediction tasks, the answer depends on future outcomes influenced by unobservable factors, where a *deterministic CoT can be problematic as it may lead the model to be overconfident* about uncertain outcomes. Instead, we propose Scenario-Aware CoT, which teaches the model to consider *multiple scenarios* before making a *probabilistic* judgment. Each CoT follows a five-phase structure:

Phase 1. **Extract**: (Same as assessment) Identify relevant quantities from the data.

Phase 2. **Compute**: Calculate price-based features (e.g., drawdown depth, recent momentum).

Phase 3. **Scenario Analysis**: Generate three scenarios:

- *Base case*: The most probable outcome given current price patterns and no external events.
- *Adverse scenario*: An external shock that could reverse the expected outcome.
- *Favorable scenario*: A positive catalyst that strengthens the expected outcome.

Phase 4. **Assessment**: Consider all scenarios to determine which outcome is most supported.

Phase 5. **Judgment**: Select the outcome most supported by the scenario analysis and input data.

Scenario templates. For each (task, answer) pair, we define domain-specific scenario templates grounded in financial knowledge. For example, *Volatility Forecast* scenarios reference volatility clustering (base), event-driven spikes (adverse), and market calming (favorable). During training, the model learns these patterns; at inference, it generates scenario reasoning adapted to the specific input, not template copies (see Section 6). Scenario templates for all tasks are provided in Appendix C.

Training setup. We fine-tune two backbones via LoRA [11]: TimeOmni-1-7B [10], an LLM trained with TS reasoning SFT and RL on Qwen2.5-7B, and Qwen2.5-7B-Instruct [23], the same base LLM without TS-specific training. As shown in Table 5, these are the best-performing models from their respective categories. We use LoRA with r=32, α=64, learning rate 5×10^{-5}, 4 epochs, batch size 1 with gradient accumulation 16, max sequence length 4096, trained on 2× NVIDIA L40S GPUs. Unless otherwise stated, FinSTaR refers to TimeOmni-1-7B fine-tuned with our CoT strategies.

5 Experiments

5.1 Baselines

We compare FinSTaR against four categories of *reasoning/language model* baselines: **(1) Language Models**: Qwen2.5-7B-Instruct [23], Llama-3.1-8B-Instruct [9], Mistral-7B-Instruct [13], Gemma-2-9B-it [21], Phi-3.5-mini [1]. **(2) TS Language Models**: TimeMQA [33]

Table 5: Results of assessment/prediction tasks across Test A/B/C. FinSTaR outperforms all baselines on 10 tasks (4 assessment, 6 prediction) under both in-domain (A) and out-of-distribution (B, C) settings. Red bold = best, blue = second best per split.

Model	Split	*Assessment*				*Prediction*						Avg.
		Draw.	Vol.	Trend	Corr.	Event	S/R	DDR	V.F.	R.P.	P.C.	
Random	—	25.0	33.3	20.0	33.3	50.0	50.0	50.0	50.0	50.0	50.0	41.2
Language Models												
Phi-3.5-mini	A	1.8	0.4	9.6	8.2	0.2	0.8	1.2	2.0	1.5	1.7	2.7
	B	3.0	1.1	12.4	5.7	0.2	1.1	1.6	0.5	1.3	1.8	2.9
	C	2.0	0.3	10.5	7.3	0.0	0.5	1.1	0.8	1.8	0.8	2.5
Mistral-7B	A	8.8	0.7	1.0	0.0	45.3	52.4	4.9	40.4	23.7	34.8	21.2
	B	4.0	1.1	0.5	0.0	37.6	46.3	2.3	40.7	24.6	36.6	19.4
	C	10.4	0.3	1.4	0.0	46.1	51.0	1.8	40.2	28.3	36.9	21.6
Gemma-2-9B	A	79.6	0.0	59.6	29.4	45.4	29.4	22.3	17.2	0.0	0.1	28.3
	B	78.9	0.0	58.4	30.9	47.4	36.9	28.0	16.4	0.1	0.0	29.7
	C	77.3	0.0	50.7	30.8	45.4	30.6	23.2	20.1	0.0	0.2	27.8
Llama-3.1-8B	A	57.2	30.3	75.7	1.4	47.3	50.2	47.3	48.0	29.9	33.6	42.1
	B	59.9	30.5	73.4	2.8	46.5	48.3	45.2	48.8	26.9	31.7	41.4
	C	57.8	30.2	76.4	2.0	46.4	50.2	46.3	45.9	31.5	35.1	42.2
Qwen2.5-7B	A	72.9	37.1	76.9	26.9	52.2	51.2	54.1	44.9	49.2	50.4	51.6
	B	71.5	37.0	73.3	30.3	53.6	48.6	58.8	46.7	**50.2**	46.6	51.7
	C	73.7	36.6	77.5	25.8	51.4	50.4	55.0	44.9	51.3	47.4	51.4
TS Language Models												
TimeMQA-Qwen	A	5.9	1.6	3.3	0.4	1.2	2.1	1.6	1.7	1.4	2.0	2.1
	B	5.5	0.8	2.9	0.6	0.7	1.7	0.8	2.1	0.7	0.8	1.7
	C	5.1	1.0	3.3	0.3	1.2	2.0	1.1	1.7	1.4	1.4	1.8
TimeMQA-Mistral	A	5.5	1.6	4.7	1.3	2.5	11.5	10.8	11.0	9.9	5.4	6.4
	B	3.5	2.8	5.4	0.9	3.0	10.1	10.9	10.8	7.0	3.9	5.8
	C	4.7	2.6	5.5	0.8	3.9	12.7	13.3	11.1	7.9	4.9	6.7
TimeMQA-Llama	A	6.9	7.5	6.6	3.3	16.0	20.4	17.6	7.2	8.7	9.7	10.4
	B	9.6	8.6	7.9	6.1	15.6	20.1	16.8	7.6	7.9	11.9	11.2
	C	10.3	6.3	5.8	6.9	16.6	18.6	15.7	8.7	9.7	10.2	10.9
TS Reasoning Models												
TimeOmni-1-7B	A	75.7	35.6	50.5	31.5	49.1	54.1	59.3	58.2	50.7	48.4	51.3
	B	75.0	35.6	53.1	31.9	51.7	50.7	57.6	55.8	48.3	49.0	50.9
	C	73.0	33.7	49.7	34.5	48.6	52.8	53.7	54.1	50.8	48.4	49.9
SFT Baselines (trained on FinTSR-Bench)												
Qwen2.5-7B (w/o CoT)	A	77.5	40.4	75.2	34.7	53.0	53.7	67.1	60.6	50.7	51.3	56.4
	B	79.6	41.6	75.9	36.2	50.0	51.1	68.7	58.2	46.9	48.1	55.6
	C	77.9	39.1	72.8	30.3	50.4	55.2	61.5	59.0	50.3	49.4	54.6
Qwen2.5-7B (w/ CoT)	A	96.4	46.6	92.3	35.4	47.6	51.1	57.2	48.7	47.3	55.0	57.8
	B	97.1	52.6	93.5	36.0	50.7	49.4	59.1	46.8	49.9	51.6	58.7
	C	96.5	45.7	91.6	38.5	48.7	51.0	56.4	49.8	45.8	53.6	57.8
Financial TSRMs (Ours)												
FinSTaR	A	**99.3**	**93.5**	**99.3**	**88.1**	**57.2**	**80.5**	**74.4**	**79.5**	**51.5**	**65.8**	**78.9**
	B	**99.6**	**92.2**	**99.5**	**89.0**	**55.6**	**80.0**	**75.0**	**77.3**	50.1	**64.3**	**78.3**
	C	**99.0**	**92.7**	**99.4**	**87.9**	**53.5**	**79.5**	**72.9**	**78.8**	**53.0**	**64.6**	**78.1**

with Qwen-2.5-7B/Mistral-7B/Llama-3-8B backbones. **(3) TS Reasoning Models**: TimeOmni-1-7B [10][1], which serves as FinSTaR's base model before fine-tuning. **(4) SFT Baselines**: To ensure fair comparison, we also fine-tune Qwen2.5-7B on FinTSR-Bench with both answer-only (w/o CoT) and full CoT supervision, providing matched baselines that isolate the contribution of TS reasoning pre-training and our CoT design. Additionally, we compare against *forecasting-based* approaches (statistical and DL models) on prediction tasks in Section 6. All LLM baselines in (1)–(3) are evaluated zero-shot, following the standard protocol in TSRM literature [10, 18]. This tests inherent reasoning ability rather than in-context learning from few-shot exemplars. We note that models with high success rates but low accuracy confirm that output formatting is not the primary bottleneck, as shown in Table 8.

[1]As of this work, TimeOmni-1 is the only TSRM that publicly releases model weights at the 7B scale. The remaining TSRMs either lack public code/weights or are available only at larger scales (e.g., Thoth, 30B).

5.2 Main Results

Table 5 presents the comparison across all three test splits. FinSTaR achieves **78.9%** overall accuracy on Test A, outperforming the best zero-shot LLM (Qwen2.5-7B, 51.6%) by **+27.3%**, the best SFT baseline (Qwen2.5-7B w/ CoT, 57.8%) by **+21.1%**, and the best TSRM (TimeOmni-1-7B, 51.3%) by **+27.6%**, with consistent performance across Test B (78.3%) and Test C (78.1%).

Key observations. 1) Assessment tasks reach near-perfect accuracy, as Compute-in-CoT makes the computation explicit and verifiable. 2) On prediction tasks, FinSTaR consistently outperforms all baselines, with the largest gains on Volatility Forecast (+21.3% over the best zero-shot baseline) and Support/Resistance (+26.4%). 3) TimeMQA models, despite being fine-tuned on TS QA, perform *worse* than general LLMs, indicating that non-financial TS training *does not transfer* to financial reasoning. 4) Many models exhibit low success rates on our benchmark as shown in Table 8, failing to produce valid output format, consistent with previous work [10].

Table 6: Ablation study w/ and w/o CoT. CoT consistently improves accuracy across all 10 tasks, with the largest gains on assessment tasks where computation becomes explicit.

	Assessment				*Prediction*						Avg.
	Draw.	Vol.	Trend	Corr.	Event	S/R	DDR	V.F.	R.P.	P.C.	
w/o CoT	78.3	37.2	68.3	32.1	49.0	56.3	66.8	69.1	46.9	54.6	55.9
w/ CoT	**99.3**	**93.5**	**99.3**	**88.1**	**57.2**	**80.5**	**74.4**	**79.5**	**51.5**	**65.8**	**78.9**
Δ	+21.0	+56.3	+31.0	+56.0	+8.2	+24.2	+7.6	+10.4	+4.6	+11.2	+23.0

Table 7: Effect of Scenario-Aware CoT. Scenario-Aware CoT improves both assessment and prediction over (deterministic) standard CoT on Test A. As expected, *the gain is larger on prediction tasks*, for which the scenario-based reasoning was specifically designed to *handle uncertainty*.

CoT	*Assessment*					*Prediction*						
	Draw.	Vol.	Trend	Corr.	Avg.	Event	S/R	DDR	V.F.	R.P.	P.C.	Avg.
Standard	99.1	92.7	**99.3**	86.9	94.5	51.0	80.0	74.3	69.3	51.1	65.4	65.2
Scenario-Aware	**99.3**	**93.5**	**99.3**	**88.1**	**95.0**	**57.2**	**80.5**	**74.4**	**79.5**	**51.5**	**65.8**	**68.2**
Δ	+0.2	+0.8	+0.0	+1.2	+0.5	+6.2	+0.5	+0.1	+10.2	+0.4	+0.4	+3.0

(a) LOCO ablation. **(b) Solo vs. joint training.**

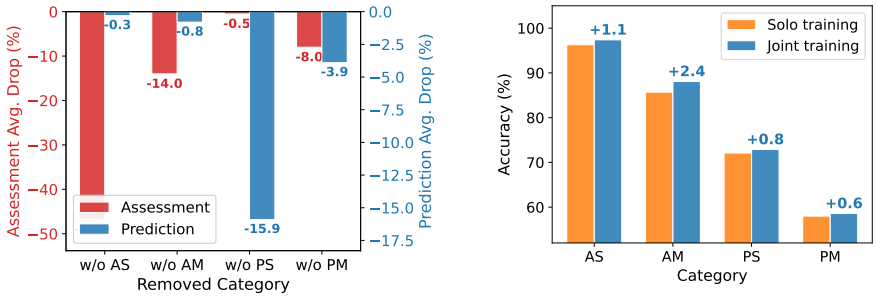


Figure 4: Taxonomy dependency analysis. (a) Each category is essential to its own tasks: removing AS causes the largest assessment drop and removing PS causes the largest prediction drop. (b) Joint training across all four categories consistently outperforms solo training, indicating that the categories are complementary.

Table 8: Success rate on Test A.

Model	SR	Acc.
Language Models		
Phi-3.5-mini	4.8	2.7
Mistral-7B	41.6	21.2
Gemma-2-9B	51.8	28.3
Llama-3.1-8B	80.3	42.1
Qwen2.5-7B	99.9	51.6
TS Language Models		
TimeMQA-Qwen	4.8	2.1
TimeMQA-Mistral	16.3	6.4
TimeMQA-Llama	25.4	10.4
TS Reasoning Models		
TimeOmni-1-7B	98.8	51.3
Ours		
FinSTaR	**100.0**	**78.9**

Note that the three test splits, which vary both stock universe and time period, implicitly *introduce input distribution shifts* (e.g., different price ranges, volatility regimes) spanning different macro regimes (Test B including COVID crash and bull markets; Test A/C including rate hikes and market volatility), and FinSTaR's consistent performance across all splits serves as a natural robustness test.

5.3 Ablation: Value of Reasoning

Table 6 compares w/o CoT vs. w/ CoT on TimeOmni-1-7B. CoT provides the largest gains on assessment tasks where computation becomes explicit: Correlation (+56.0), Volatility (+56.3), and Trend (+31.0). On prediction tasks, gains are moderate but consistent: Support/Resistance (+24.2), Volatility Forecast (+10.4), and Pair Convergence (+11.2). Relative Performance shows the smallest gain (+4.6), reflecting the inherent difficulty of stock return prediction.

5.4 Scenario-Aware CoT Effect

To isolate the effect of Scenario-Aware CoT, we compare two CoT configurations for prediction tasks: (deterministic) standard CoT vs. (probabilistic) Scenario-Aware CoT, with standard CoT fixed for assessment tasks in both. As shown in Table 7, assessment accuracy is nearly identical between the two. On prediction tasks, Scenario-Aware CoT achieves +2.9% average improvement, with the largest gains on Volatility Forecast (+10.2) and Event Response (+6.2). The overall average (+2.0%) is diluted by assessment tasks where both configs use identical CoT ($\Delta \approx 0$ by design). These results confirm that *scenario reasoning consistently improves performance on uncertain outcomes.*

6 Analysis

Capability dependencies. We study how the *four capability categories* — **AS** (Assessment-Single), **AM** (Assessment-Multi), **PS** (Prediction-Single), **PM** (Prediction-Multi) — interact during training via two complementary analyses on TimeOmni-1-7B on Test A. First, we conduct a *Leave-One-Category-Out* (LOCO) ablation: remove one category's tasks from training, retrain, and evaluate all ten tasks. As shown in Figure 4(a), each category is *essential to its own tasks* — removing AS causes the largest assessment drop and

Table 9: Leave-One-Category-Out (LOCO) ablation study. Each row removes one capability category from training and evaluates on all 10 tasks. Together with Table 10, the results reveal the dependencies among task categories and demonstrate the *effectiveness of joint training across all four capabilities.*

	Assessment				*Prediction*						Avg.
	Draw.	Vol.	Trend	Corr.	Event	S/R	DDR	V.F.	R.P.	P.C.	
Full	**99.3**	**93.5**	99.3	**88.1**	**57.2**	**80.5**	74.4	**79.5**	**51.5**	**65.8**	**78.9**
w/o AS	43.2	32.2	29.2	88.0	56.8	79.8	**75.3**	78.8	**51.5**	65.4	60.0
w/o AM	98.8	92.3	99.2	33.4	53.9	80.4	75.2	78.9	**51.5**	64.4	72.8
w/o PS	99.2	90.6	**99.4**	88.0	47.7	48.1	47.4	55.1	**51.5**	64.1	69.1
w/o PM	98.9	92.3	67.5	**88.1**	57.1	80.4	74.8	76.6	47.8	49.2	73.3

Table 10: Solo vs. joint training.

Cat.	Task	Solo	Joint	Δ
AS	Draw.	98.8	**99.3**	+0.5
	Vol.	91.1	**93.5**	+2.4
	Trend	99.1	**99.3**	+0.2
AM	Corr.	85.7	**88.1**	+2.4
PS	Event	55.9	**57.2**	+1.3
	S/R	79.5	**80.5**	+1.0
	DDR	**74.4**	**74.4**	+0.0
	V.F.	78.4	**79.5**	+1.1
PM	R.P.	**51.5**	**51.5**	+0.0
	P.C.	64.5	**65.8**	+1.3

Table 11: Forecasting baselines on prediction. Values are (Test A, Test B, Test C). FinSTaR outperforms forecasting baselines, with the largest gains on tasks requiring scenario reasoning.

Method	Event	S/R	DDR	V.F.	R.P.	P.C.	Avg.
Statistical Forecasting							
Last Value	(42.7, 44.5, 46.4)	(61.7, 62.1, 59.6)	(49.9, 51.2, 49.3)	(48.0, 53.6, 51.9)	(48.5, 49.9, 47.0)	(50.3, 48.3, 49.1)	(50.2, 51.6, 50.5)
ETS	(49.2, 52.0, 50.3)	(62.1, 59.5, 58.9)	(46.3, 47.6, 49.9)	(48.0, 53.6, 51.9)	(49.5, 53.9, 50.2)	(54.6, 56.5, 52.3)	(51.6, 53.8, 52.2)
MA	(51.2, 52.7, 52.1)	(60.1, 57.2, 57.8)	(45.1, 46.1, 46.3)	(50.6, 56.8, 53.9)	(50.8, 53.2, 51.3)	(59.0, 57.6, 56.5)	(52.8, 53.9, 53.0)
Momentum	(49.5, 47.2, 50.4)	(68.9, 68.8, 68.4)	(57.1, 55.6, 57.6)	(48.0, 53.6, 51.9)	(49.7, 45.9, 49.0)	(50.5, 49.5, 52.3)	(53.9, 53.4, 54.9)
Drift	(52.9, 50.7, 50.4)	(71.0, 72.5, 68.1)	(63.5, 66.0, 64.1)	(48.0, 53.6, 51.9)	(48.4, 47.8, 50.4)	(50.3, 48.3, 49.1)	(55.7, 56.5, 55.7)
DL Forecasting							
TFT	(52.8, 53.0, 50.6)	(61.6, 60.2, 58.1)	(36.3, 38.9, 39.9)	(48.0, 53.7, 51.9)	(50.4, 49.2, **54.3**)	(52.2, 50.9, 50.4)	(50.2, 51.0, 50.9)
DeepAR	(49.1, 53.1, 51.2)	(64.0, 61.4, 60.6)	(50.4, 50.3, 51.0)	(48.2, 54.2, 52.0)	(48.4, 53.7, 49.9)	(56.4, 55.9, 56.2)	(52.8, 54.8, 53.5)
DLinear	(47.9, 48.7, 52.0)	(63.8, 62.1, 58.5)	(64.9, 66.9, 63.0)	(58.2, 55.3, 57.7)	(46.4, 52.5, 50.2)	(55.0, 56.8, 54.8)	(56.0, 57.0, 56.0)
TiDE	(54.1, 54.4, 53.0)	(57.4, 54.6, 56.7)	(58.1, 57.2, 56.7)	(58.9, 55.6, 58.0)	(50.5, **54.1**, 50.3)	(56.2, 57.5, 57.0)	(55.9, 55.6, 55.3)
PatchTST	(48.6, 49.2, 49.6)	(80.4, 79.7, 79.1)	(49.3, 48.7, 49.1)	(69.5, 64.6, 68.9)	(48.2, 48.6, 48.5)	(48.4, 44.9, 47.7)	(57.4, 56.0, 57.1)
Chronos-1	(53.4, 54.9, 48.9)	(69.4, 71.0, 66.5)	(50.2, 53.9, 50.1)	(47.9, 54.2, 52.0)	(50.1, 48.0, 49.8)	(53.0, 51.0, 53.0)	(54.0, 55.5, 53.4)
Chronos-2	(53.2, 53.0, 51.1)	(67.9, 69.5, 64.4)	(45.3, 47.0, 46.3)	(48.0, 53.6, 51.9)	(51.4, 49.1, 52.4)	(53.2, 48.2, 51.3)	(53.2, 53.4, 52.9)
Financial TSRM (Ours)							
FinSTaR	(**57.2**, **55.6**, **53.5**)	(**80.5**, **80.0**, **79.5**)	(**74.4**, **75.0**, **72.9**)	(**79.5**, **77.3**, **78.8**)	(**51.5**, 50.1, 53.0)	(**65.8**, **64.3**, **64.6**)	(**68.2**, **67.0**, **67.0**)

removing PS causes the largest prediction drop — while removals do not catastrophically harm the other axis. Second, Figure 4(b) compares solo training (each category alone) with joint training (all four together): joint training *consistently outperforms* solo training across all categories. Together, these results show that the four categories are *complementary and mutually reinforcing through joint training*, rather than redundant or interfering. Full results are shown in Tables 9 and 10.

Success rate. Table 8 reports the success rate (SR), the percentage of outputs containing a valid answer tag. Most baselines fail to follow the required format: TimeMQA variants and Phi-3.5-mini achieve SR < 25%, rendering their accuracy unreliable. In contrast, FinSTaR achieves 100% SR, confirming that structured CoT supervision also teaches reliable output formatting.

Prediction tasks: TSRM vs. TS forecasting models. While TSRMs can handle both assessment and prediction through reasoning, traditional TS forecasting models can also solve prediction tasks without reasoning, via a *forecast-then-classify* pipeline (predict future prices numerically, then convert to classification labels). We compare FinSTaR against such forecasting models: statistical methods (Last Value, MA, ETS, Drift, Momentum) and DL models (DLinear [32], Chronos-1/2 [2], DeepAR [22], PatchTST [19], TFT [15], TiDE [8]). As shown in Table 11, FinSTaR outperforms the forecasting models on prediction tasks. Details of forecasting models are discussed in Appendix B.

Backbone sensitivity. Table 12 compares the effect of CoT across two backbones: TimeOmni-1-7B (trained with TS reasoning SFT + RL) and Qwen2.5-7B (general-purpose LLM without TS-specific training).[2] CoT improves assessment accuracy for both backbones, since Compute-in-CoT involves explicit arithmetic that even general-purpose LLMs can perform. However, for prediction tasks, CoT improves TimeOmni-1 (+11.3%) but *hurts* Qwen (−5.0%). We attribute this to the fact that Scenario-Aware CoT requires reasoning about price dynamics, which demands prior TS reasoning ability. Without TS reasoning training, the general-purpose LLM generates plausible-sounding but ungrounded scenarios, introducing noise rather than useful reasoning signal. This suggests that *Scenario-Aware CoT is effective only when the backbone already possesses TS reasoning ability*, highlighting the synergy between TS reasoning pre-training and structured CoT supervision.

[2]Both Qwen2.5-7B rows (w/o CoT and w/ CoT) are fine-tuned on the same FinTSR-Bench data via LoRA, serving as SFT baselines that isolate the effect of TS reasoning pre-training from our CoT strategies.

Table 12: Backbone sensitivity. Scenario-Aware CoT improves prediction only with a TS reasoning-trained backbone (TimeOmni-1, +11.3%), while it hurts the general-purpose LLM (Qwen, −5.0%), suggeting that *prior TS reasoning ability is necessary* for effective scenario-based reasoning.

Backbone	Config	Assessment Avg.	Prediction Avg.	Overall
Qwen2.5-7B (w/o TS Reasoning SFT)	w/o CoT	57.0	**56.1**	56.4
	w/ CoT	**67.7**	51.1	**57.8**
	Δ	+10.7	−5.0	+1.3
TimeOmni-1-7B (w/ TS Reasoning SFT)	w/o CoT	54.3	56.9	55.9
	w/ CoT	**95.0**	**68.2**	**78.9**
	Δ	+40.7	+11.3	+23.1

Table 13: Effect of training set size. FinSTaR performance with 350, 875, 1750, and 3500 training samples per task. With only 10% of the full data (350 samples/task), FinSTaR already achieves 71.7%, indicating that *structured CoT strategies are sample-efficient.*

Method	# Samples	*Assessment*				*Prediction*						Avg.
		Draw.	Vol.	Trend	Corr.	Event	S/R	DDR	V.F.	R.P.	P.C.	
TimeOmni-1-7B	3500 (100%)	75.7	35.6	50.5	31.5	49.1	54.1	59.3	58.2	50.7	48.4	51.3
Qwen2.5-7B (w/o CoT)	3500 (100%)	77.5	40.4	75.2	34.7	53.0	53.7	67.1	60.6	50.7	51.3	56.4
Qwen2.5-7B (w/ CoT)	3500 (100%)	96.4	46.6	92.3	35.4	47.6	51.1	57.2	48.7	47.3	55.0	57.8
FinSTaR	350 (10%)	98.1	80.4	98.0	71.6	53.9	54.8	70.6	73.9	51.4	63.8	71.7
	875 (25%)	98.0	87.3	99.0	83.1	57.4	53.9	73.6	78.2	51.5	64.5	74.7
	1750 (50%)	99.1	90.0	82.2	85.5	54.5	70.9	74.7	78.7	51.5	65.0	75.2
	3500 (100%)	99.3	93.5	99.3	88.1	57.2	80.5	74.4	79.5	51.5	65.8	78.9

Effect of training set size. We evaluate FinSTaR's performance with a reduced training data size. As shown in Table 13, FinSTaR already achieves 71.7% overall accuracy with only 350 samples/task (10% of full). Performance improves to 74.7% at 875 and 75.2% at 1750 samples, with diminishing returns beyond that. This indicates that our Compute-in-CoT and Scenario-Aware CoT strategies are *data-efficient*, requiring relatively few examples to learn the underlying reasoning patterns.

CoT quality verification. We verify that FinSTaR generates meaningful reasoning rather than template copies along four dimensions, with details provided in Appendix D:

- **Format compliance.** Over 99% of outputs contain valid `<think>` and `<answer>` tags.
- **Arithmetic faithfulness.** Since each assessment label is uniquely determined by a numerical computation, the near-perfect accuracy implies faithful intermediate reasoning. Manual inspection confirms that computed values align with ground truth.
- **Scenario diversity.** Across 100 sampled prediction outputs, 85+ unique scenario descriptions per task confirm that the model *generates varied reasoning, not memorized templates.* While scenario templates are defined per (task, answer) pair during training, this does not constitute label leakage: the model generates all three scenarios *before* selecting an answer, and the template serves as a reasoning scaffold rather than an answer key.
- **Output length.** Prediction CoT averages ∼249 tokens, approximately 2.4× longer than assessment CoT (∼104 tokens), reflecting the additional scenario analysis steps.

7 Conclusion

We introduce FinSTaR, *the first Financial TSRM*, which separates financial TS reasoning into assessment and prediction, applying Compute-in-CoT and Scenario-Aware CoT respectively. Our general 2×2 capability taxonomy, instantiated as ten financial reasoning tasks, reveals that the four categories are complementary and mutually reinforcing through joint training, and our scenario-based reasoning consistently outperforms standard CoT on prediction tasks. FinSTaR achieves 78.9% overall accuracy, substantially outperforming 15+ baselines. We will publicly release FinTSR-Bench, FinSTaR model checkpoints, and all evaluation code under an open-source license upon publication.

Limitations and future work. Our benchmark covers only U.S. equities and 7B-parameter models, and prediction tasks exhibit a performance ceiling reflecting fundamental market efficiency. Future directions include extending to other asset classes and additional inputs (e.g., news sentiment, fundamentals), as well as leveraging a TS encoder for multimodal reasoning.

Impact Statement

This work advances time series reasoning by introducing a general capability taxonomy and instantiating it in the financial domain through FinTSR-Bench and FinSTaR. The positive societal impact includes enabling more interpretable and transparent financial analysis—our structured chain-of-thought framework produces grounded, step-by-step reasoning that is auditable by analysts and regulators, in contrast to opaque black-box forecasters. By distinguishing deterministic assessment from stochastic prediction, our approach also promotes epistemically honest modeling: FinSTaR explicitly acknowledges uncertainty in prediction tasks through scenario-based reasoning rather than producing overconfident point predictions. However, we caution against unintended misuse: the model is trained on historical S&P 500 data and should not be deployed as an autonomous trading system or investment advisor, as prediction tasks are inherently uncertain (reflecting market efficiency) and outputs may reflect biases in the training distribution. We recommend using FinSTaR as a decision-support tool that augments, rather than replaces, human financial judgment, and we emphasize that model outputs should be validated before any consequential financial decisions.

References

[1] Marah Abdin et al. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv preprint arXiv:2404.14219* (2024).

[2] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, et al. 2024. Chronos: Learning the Language of Time Series. *Transactions on Machine Learning Research* (2024).

[3] Wei Chen et al. 2026. SenTSR-Bench: Thinking with Injected Knowledge for Time-Series Reasoning. *arXiv preprint arXiv:2602.19455* (2026).

[4] Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023. Program of Thoughts Prompting: Disentangling Computation from Reasoning for Numerical Reasoning Tasks. In *TMLR*.

[5] Yihan Chen et al. 2025. Bridging Time and Linguistics: LLMs as Time Series Analyzer through Symbolization and Segmentation. In *NeurIPS*.

[6] Yue Chen et al. 2025. Visual-Temporal Analysis: Multimodal Reasoning over Financial Time Series and Charts. *arXiv preprint* (2025).

[7] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

[8] Abhimanyu Das, Weihao Kong, Andrew Leber, Rajat Mathews, and Rajat Sen. 2023. Long-term Forecasting with TiDE: Time-series Dense Encoder. In *TMLR*.

[9] Aaron Grattafiori et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783* (2024).

[10] Tong Guan, Zijie Meng, Dianqi Li, Shiyu Wang, Chao-Han Huck Yang, Qingsong Wen, Zuozhu Liu, Sabato Marco Siniscalchi, Ming Jin, and Shirui Pan. 2026. Timeomni-1: Incentivizing complex reasoning with time series in large language models. *ICLR* (2026).

[11] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*.

[12] Rob J Hyndman and George Athanasopoulos. 2021. *Forecasting: Principles and Practice* (3rd ed.). OTexts.

[13] Albert Q Jiang et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).

[14] Jian Li et al. 2025. Time-R1: Towards Temporal Reasoning with Reinforcement Learning in LLMs. *arXiv preprint* (2025).

[15] Bryan Lim, Sercan Ö Arık, Nicolas Loeff, and Tomas Pfister. 2021. Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting. In *International Journal of Forecasting*.

[16] Yifan Liu et al. 2025. FinTSB: A Comprehensive and Practical Benchmark for Financial Time Series Forecasting. *arXiv preprint arXiv:2502.18834* (2025).

[17] Zhiming Liu et al. 2025. Reinforcement Learning for Financial Verbal Reasoning with Large Language Models. *arXiv preprint* (2025).

[18] Mike A Merrill et al. 2024. Towards Time Series Reasoning with LLMs. *arXiv preprint arXiv:2409.11376* (2024).

[19] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *ICLR*.

[20] Jiwon Park et al. 2026. PATRA: Pattern-Aware Alignment and Balanced Reasoning for Time Series Question Answering. *arXiv preprint arXiv:2602.23161* (2026).

[21] Morgane Riviere et al. 2024. Gemma 2: Improving Open Language Models at a Practical Size. *arXiv preprint arXiv:2408.00118* (2024).

[22] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. 2020. DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks. In *International Journal of Forecasting*.

[23] Qwen Team. 2025. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115* (2025).

[24] Shuangge Wang et al. 2024. CausalStock: Deep End-to-end Causal Discovery for News-driven Multi-stock Movement Prediction. In *NeurIPS*.

[25] Xinlei Wang et al. 2024. From News to Forecast: Integrating Event Analysis in LLM-Based Time Series Forecasting with Reflection. In *NeurIPS*.

[26] Yifan Wang et al. 2025. COUNTS: Benchmarking LLM Numerical Reasoning with Verifiable Rewards. *arXiv preprint arXiv:2501.00000* (2025).

[27] Yanlong Wang, Jian Xu, Fei Ma, Hongkang Zhang, Hang Yu, Tiantian Gao, Yu Wang, Haochen You, Shao-Lun Huang, Danny Dongning Sun, et al. 2025. FinZero: Launching Multi-modal Financial Time Series Forecast with Large Reasoning Model. *arXiv preprint arXiv:2509.08742* (2025).

[28] Zhihao Wang et al. 2025. MTBench: Multi-Task Temporal Reasoning Benchmark. *arXiv preprint* (2025).

[29] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

[30] Qianqian Xie et al. 2024. FinBen: A Holistic Financial Benchmark for Large Language Models. In *NeurIPS Datasets and Benchmarks*.

[31] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. STaR: Bootstrapping Reasoning With Reasoning. In *NeurIPS*.

[32] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. 2023. Are Transformers Effective for Time Series Forecasting?. In *AAAI*.

[33] Yi Zhang et al. 2025. Time-MQA: Time Series Multi-Task Question Answering with Context Enhancement. *ACL* (2025).

[34] Yifan Zhang et al. 2025. TimeART: Tool-Augmented Reasoning for Time Series. *arXiv preprint* (2025).

[35] Yifan Zhang et al. 2026. Thoth: Mid-Training Bridges LLMs to Time Series Understanding. *arXiv preprint arXiv:2603.01042* (2026).

A Dataset Construction

Table A.1 lists the QA generation parameters used for constructing FinTSR-Bench. All tasks use a 120-day input window of daily closing prices. Assessment tasks evaluate the current state and have no prediction horizon. For prediction tasks, Event Response and Support/Resistance use a 10-day forward horizon, while Drawdown Recovery, Volatility Forecast, Relative Performance, and Pair Convergence use a 20-day forward horizon. Table A.2 shows the training set class distribution, confirming that all tasks are approximately balanced via undersampling.

Table A.1: QA generation parameters.

Parameter	Value
Window length	120 trading days
Samples per task (raw)	100,000
Cap per task (train)	3,500
Cap per task (test)	1,000
Correlation thresholds	pos > 0.30, neg < -0.10
Event threshold	$\lvert z\rvert > 2.5$
Rel. Perf. margin	5%
Rel. Perf. forward window	20 days
S/R breakout margin	3%
Drawdown Recovery min drawdown	5%
Volatility Forecast change threshold	25%
Pair Convergence spread margin	3%

Table A.2: Training set class distribution. All tasks are approximately class-balanced via undersampling.

Task	Classes	Distribution (%)
Drawdown	4	A: 24.2, B: 24.3, C: 26.0, D: 25.5
Volatility Regime	3	A: 33.5, B: 33.1, C: 33.4
Trend Direction	5	A: 20.2, B: 20.4, C: 19.6, D: 21.0, E: 18.7
Correlation	3	A: 32.9, B: 34.2, C: 32.9
Event Response	2	A: 49.0, B: 51.0
Support/Resistance	2	A: 50.1, B: 49.9
Drawdown Recovery	2	A: 49.0, B: 51.0
Volatility Forecast	2	A: 48.9, B: 51.1
Relative Performance	2	A: 51.4, B: 48.6
Pair Convergence	2	A: 50.5, B: 49.5

Label construction algorithms. We detail the labeling procedure for each prediction task to ensure reproducibility.

- **Event Response.** We compute daily returns and their z-scores over the 120-day window. A trading day with $\lvert z\rvert > 2.5$ is flagged as an "event." The label is determined by comparing the 10-day post-event return direction with the event-day return: same direction → Persistence, opposite → Mean-reversion.
- **Support/Resistance.** We identify support (resistance) as the minimum (maximum) closing price over a rolling 60-day lookback window. If the current price is within 5% of a key level, the sample is included. The label is determined by whether the price moves beyond the level by >3% (Breakout) or stays within it (Bounce) over the next 10 days.
- **Drawdown Recovery.** We compute peak price as the running maximum over the window. Stocks with drawdown >5% from peak are included. The label is Recovery if the price moves >3% toward the peak over the next 20 days, and Deepens otherwise.
- **Volatility Forecast.** We compute the ratio of 20-day to 120-day return standard deviation. The label is Increase if the 20-day volatility computed at t+20 exceeds the current 20-day volatility by >25%, and Decrease otherwise.
- **Relative Performance.** Given two stocks, the label is determined by which stock achieves a higher cumulative return over the next 20 days. Pairs with <5% return difference are excluded to avoid ambiguous labels.
- **Pair Convergence.** We compute the normalized spread as $(P_A - P_B)/(P_A + P_B)$, where P_A and P_B are the closing prices. The label is Convergence if $\lvert\text{spread}_{t+20}\rvert < \lvert\text{spread}_t\rvert$, and Divergence otherwise. Pairs with spread <3% are excluded.

To avoid overlapping-window leakage, each task uses a random seed for sample selection, and we enforce a minimum gap of 20 trading days between consecutive samples from the same stock.

B Forecasting Baseline Protocol

All forecasting baselines use the same 120-day input window as FinSTaR and predict over the task-specific horizon (10 or 20 days). No exogenous covariates are used; all models receive only univariate closing prices. Predicted price trajectories are converted to classification labels using the same thresholds as the QA generation (e.g., direction of predicted return for Event Response, predicted volatility ratio for Volatility Forecast).

Statistical methods (applied without hyperparameter tuning):

- **Last Value.** Predicts the last observed value for all future steps (naive baseline).
- **Moving Average (MA).** Predicts the mean of the last k observations (k=20).
- **Exponential Smoothing (ETS) [12].** Weighted average with exponentially decaying weights; captures level and trend.
- **Drift.** Extrapolates a linear trend from the first to the last observed value.
- **Momentum.** Predicts continuation of the recent return direction over the past 20 days.

Deep learning models (trained per-task using AutoGluon TimeSeries with default hyperparameters and early stopping):

- **DLinear [32].** Decomposes TS into trend and remainder via moving average, then applies linear layers.
- **DeepAR [22].** Autoregressive RNN that produces probabilistic forecasts via learned likelihood parameters.
- **PatchTST [19].** Patch-based Transformer that segments TS into subseries-level patches for efficient attention.
- **TFT [15].** Temporal Fusion Transformer with variable selection, gating, and multi-horizon attention.
- **TiDE [8].** MLP-based encoder-decoder with dense residual connections for long-term forecasting.

Pre-trained forecasting models (zero-shot, no fine-tuning):

- **Chronos-1 [2].** Tokenizes TS values into discrete bins and uses a T5-based language model for forecasting.
- **Chronos-2 [2].** Updated version with improved tokenization and larger pre-training corpus.

C Scenario Templates

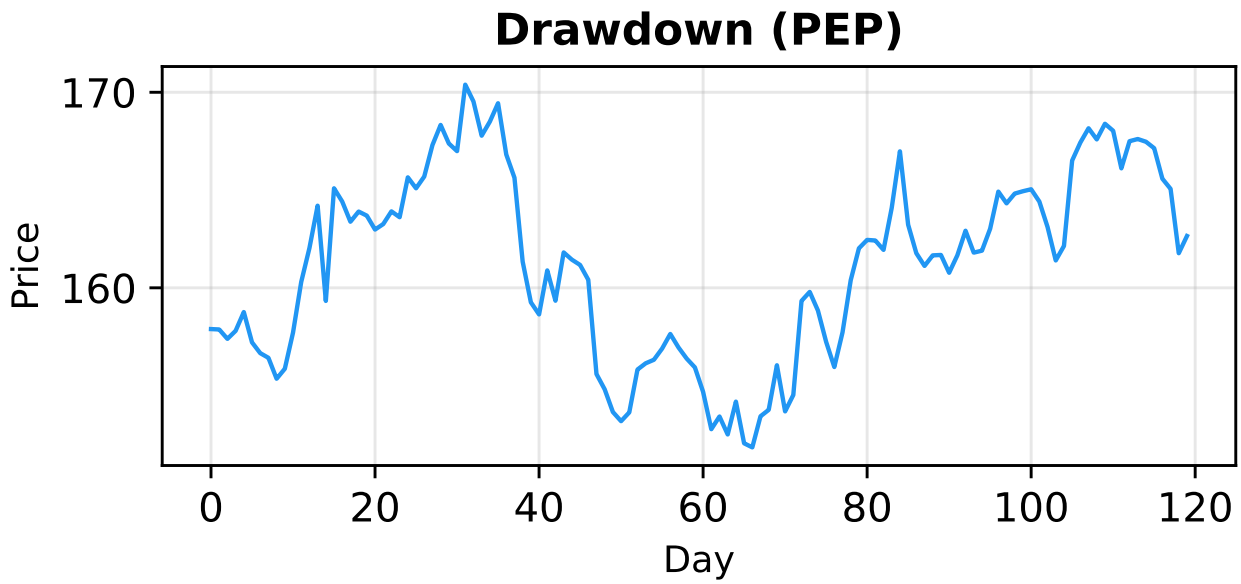
Table D.1 lists the scenario templates used in Scenario-Aware CoT for all six prediction tasks. For each task and answer, we define three scenarios (base, adverse, favorable) grounded in financial domain knowledge. Assessment tasks (Drawdown, Volatility Regime, Trend Direction, Correlation) use deterministic Compute-in-CoT and do not require scenario templates.

D Full Task Examples

We present actual model input, generated CoT reasoning, and output for all ten tasks. Each example shows a correctly predicted sample from Test A, with the corresponding price series visualization.

Table D.1: Scenario templates for prediction tasks. Each (task, answer) pair has three domain-specific scenarios.

Task	Answer	Base Case	Adverse	Favorable
Event Response	Mean-reversion	Extreme move partially reverses as market recognizes overreaction	Secondary catalyst reinforces shock direction	Positive news accelerates recovery
	Persistence	Shock reflects genuine new information being priced in	Mean-reversion forces prove it was an overreaction	Additional confirming news amplifies the move
Support/Resistance	Breakout	Price breaks through with sustained momentum	False breakout traps traders, price reverses	High volume accelerates move beyond level
	Bounce	Key level holds as institutional orders defend it	Sustained pressure overwhelms the level	Strong bounce triggers significant reversal
Drawdown Recovery	Recovery	Drawdown stabilizes, value buyers accumulate	Cause worsens (earnings revision, guidance cut)	Positive catalyst triggers V-shaped recovery
	Further decline	Selling pressure continues, drawdown deepens	Surprise positive development reverses decline	Risk-off sentiment compounds weakness
Volatility Forecast	Increase	Volatility clustering continues, uncertainty persists	Uncertainty resolves, volatility collapses	New uncertainty compounds existing volatility
	Decrease	Volatility mean-reverts as shock dissipates	Unexpected event triggers new spike	Broad market calming accelerates compression
Relative Performance	Stock A	Momentum trends persist, Stock A continues trajectory	Stock A headwind, Stock B catalyst reverses gap	Sector tailwinds boost Stock A's advantage
	Stock B	Stock B strengthening continues, Stock A fades	Stock A recovers, re-establishing outperformance	Stock B catalyst (upgrade, M&A) widens lead
Pair Convergence	Converge	Spread narrows as co-movement reasserts	Structural change sustains divergence	Common catalyst forces rapid convergence
	Diverge	Different fundamentals sustain widening spread	Common catalyst narrows the spread	Company-specific event accelerates divergence



Task: Drawdown — Ticker: PEP — GT: (B) Pullback

Input: You are analyzing the stock PEP. Below are the daily closing prices for the most recent 120 trading days: [*prices shown above*]. Based on these prices, assess the current drawdown phase.
(A) At/Near Peak (<3%)
(B) Pullback (3–10%)
(C) Correction (10–20%)
(D) Severe Decline (>20%)

Output:

<think>

Step 1 — Find the peak price:
Scanning the 120-day series, the highest price is 170.39 (around day 32).

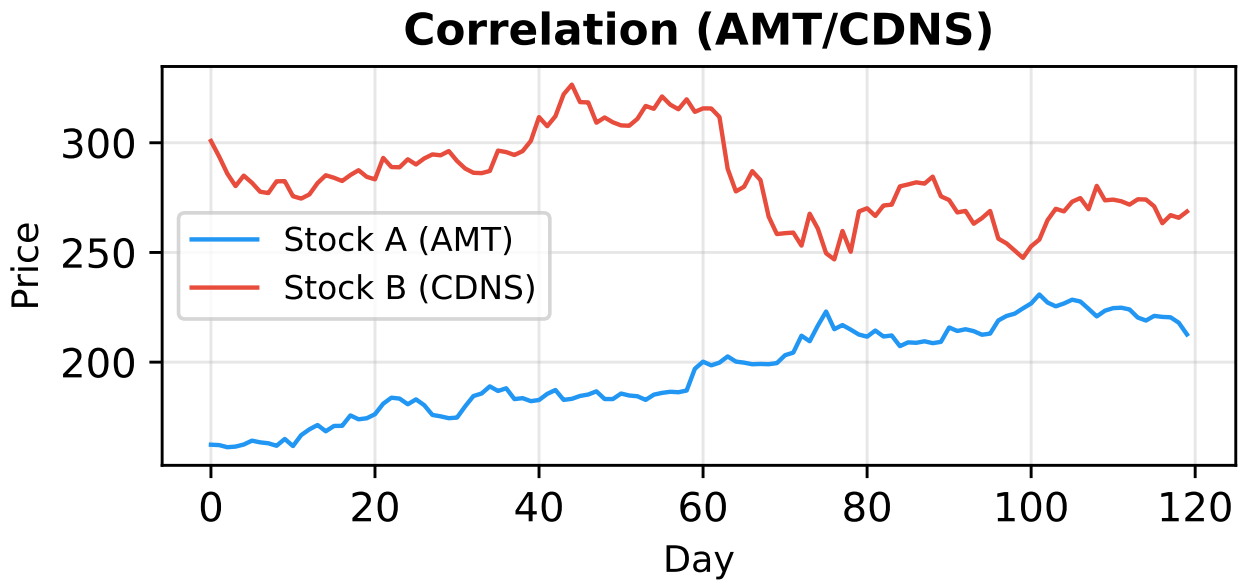
Step 2 — Current price:
The last price in the series is 162.65.

Step 3 — Calculate drawdown:
Drawdown = (Peak − Current) / Peak = (170.39 − 162.65) / 170.39 = 4.5%

Step 4 — Classify:
4.5% is between 3% and 10%, which corresponds to (B) Pullback.

</think>
<answer>(B)</answer>

Figure D.1: (Single-Assessment) Example of Drawdown task.



Task: Correlation — Tickers: AMT vs CDNS — GT: (B) Negative correlation

Input: You are analyzing the relationship between two stocks. Stock A (AMT) and Stock B (CDNS) daily closing prices (120 days each): [*prices shown above*]. Determine the correlation between their daily returns.
(A) Positive
(B) Negative
(C) No significant correlation

Output:

<think>

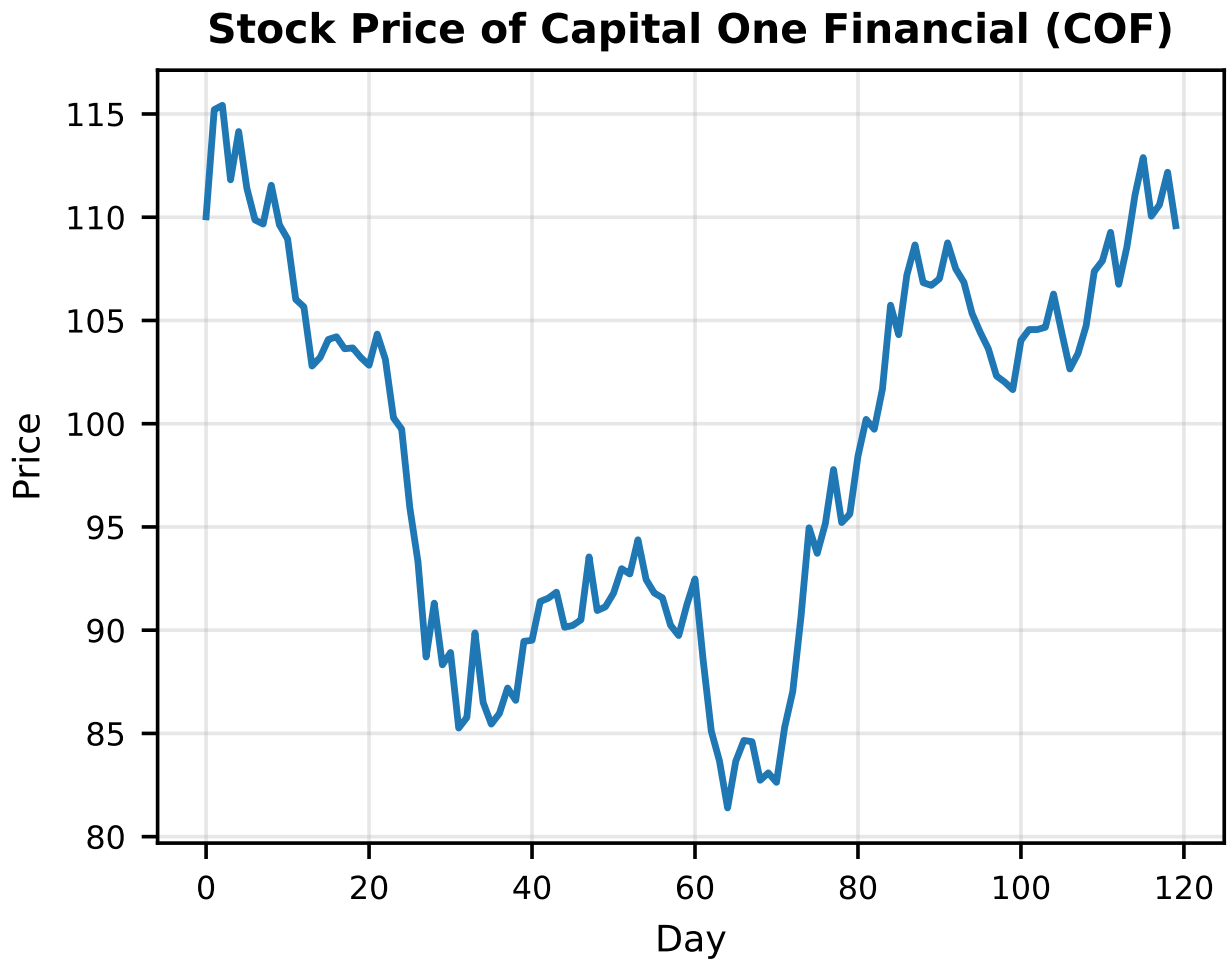
Step 1 — Compare price changes in 12 periods (~10 days each):
Days 1-10: A=+1.6%, B=−6.0%
Days 11-20: A=+7.8%, B=+3.2%
Days 21-30: A=−1.0%, B=+4.5%
Days 31-40: A=+2.3%, B=−5.1%
Days 41-50: A=−0.8%, B=+1.9%
Days 51-60: A=+3.1%, B=−2.4%
Days 61-70: A=−2.6%, B=+4.0%
Days 71-80: A=+1.5%, B=−1.8%
Days 81-90: A=−1.9%, B=+2.7%
Days 91-100: A=+0.4%, B=−0.6%
Days 101-110: A=−1.5%, B=+2.2%
Days 111-120: A=+2.4%, B=−3.1%

Step 2 — Pattern:
When A rises, B tends to fall, and when A falls, B tends to rise. The two stocks move in opposite directions across most periods.

Step 3 — negative correlation → (B).

</think>
<answer>(B)</answer>

Figure D.2: (Multi-Assessment) Example of Correlation task.



Task: Volatility Regime — Ticker: COF — GT: (B) Normal volatility

Input: You are analyzing the stock COF. Below are the daily closing prices (120 days): [*prices shown above*]. Compute the ratio of recent (last 20 days) to overall (120-day) volatility.
(A) Low (<0.6)
(B) Normal (0.6–1.6)
(C) High (>1.6)

Output:

<think>

Step 1 — Compute daily returns from prices:
e.g., (54.32→54.51): +0.35%, (54.51→54.10): −0.75%, (54.10→54.78): +1.26%
...

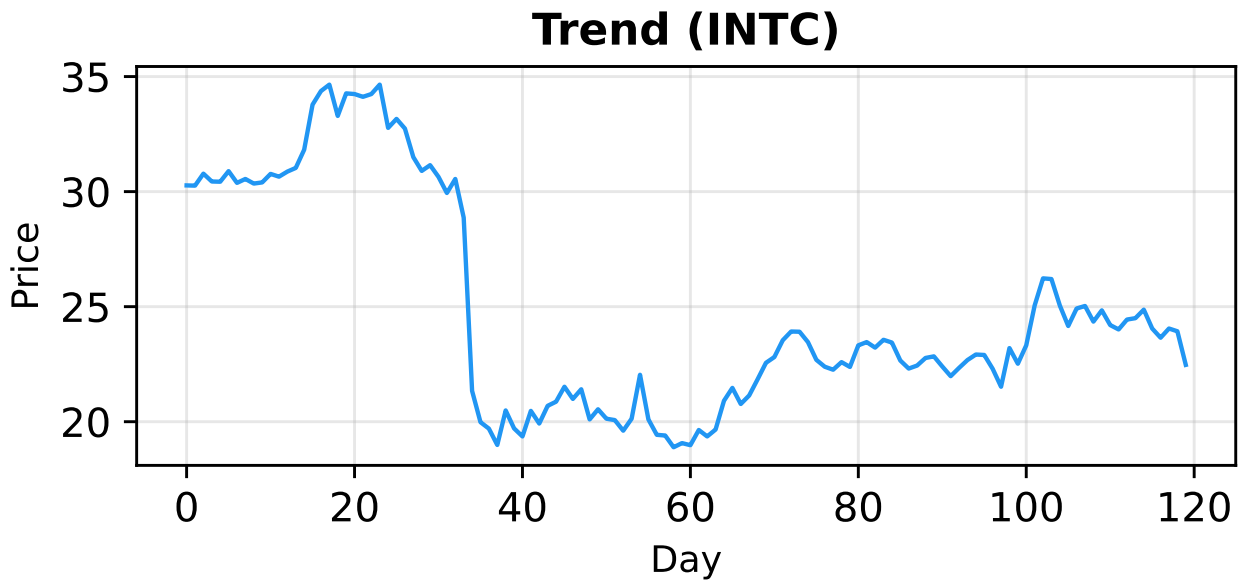
Step 2 — Overall volatility (std of all returns): 2.02%

Step 3 — Recent 20-day volatility: 1.50%

Step 4 — Ratio = 1.50 / 2.02 = 0.74
Recent volatility is close to the long-term average — normal conditions.
→ (B) normal volatility.

</think>
<answer>(B)</answer>

Figure D.3: (Single-Assessment) Example of Volatility Regime task.



Task: Trend Direction — Ticker: INTC — GT: (E) Strong Downtrend

Input: You are analyzing the stock INTC. Below are the daily closing prices (120 days): [*prices shown above*]. Classify the overall trend direction.
(A) Strong Uptrend (>20%)
(B) Mild Uptrend (5–20%)
(C) Sideways (−5% to +5%)
(D) Mild Downtrend (5–20% decline)
(E) Strong Downtrend (>20% decline)

Output:

<think>

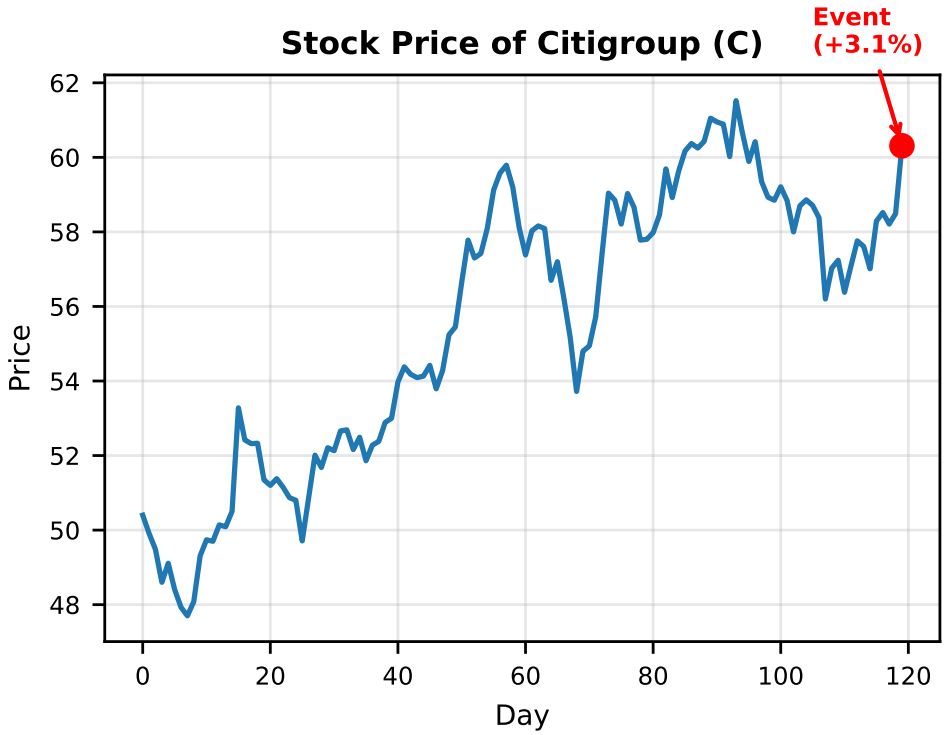
Step 1 — Read start and end prices:
Start (day 1): 30.27
End (day 120): 22.47

Step 2 — Compute cumulative return:
Return = (22.47 − 30.27) / 30.27 = −25.8%

Step 3 — Classify:
−25.8% → (E) Strong Downtrend (< −20%).

</think>
<answer>(E)</answer>

Figure D.4: (Single-Assessment) Example of Trend Direction task.



Task: Event Response — Ticker: C — GT: (B) Persistence

Input: You are analyzing the stock C. Below are the daily closing prices (120 days): [*prices shown above*]. The stock just experienced a significant positive shock. Predict the outcome over the next 10 days.
(A) Mean-reversion
(B) Persistence

Output:

<think>

Step 1 — Event: positive shock, z=2.20, return=+1.85%.

Step 2 — Pre-event trend: flat.

Step 3 — Scenario analysis:
• *Base case* (most probable — no major external events):
The positive shock reflects genuine new information being priced in. The market continues to absorb the news over the following days, sustaining the move in the same direction.

• *Adverse scenario* (external shock reverses expectation):
Mean-reversion forces kick in as profit-taking and overreaction corrections pull the price back toward pre-event levels.

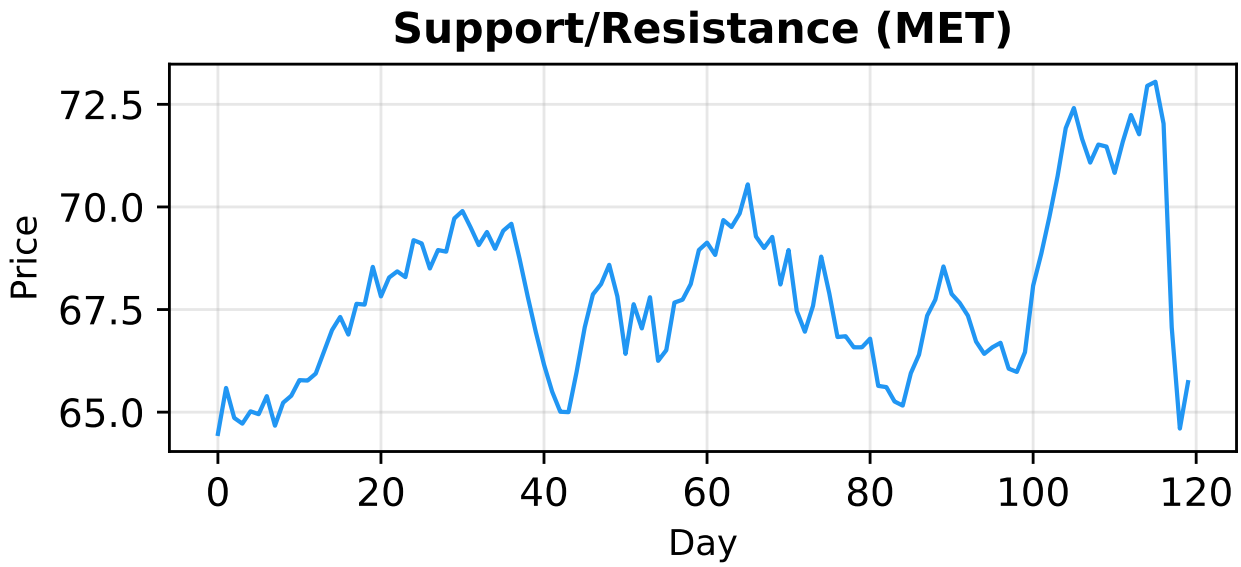
• *Favorable scenario* (catalyst strengthens expectation):
Additional confirming news (e.g., earnings beat, upgrade) amplifies the move and accelerates persistence.

Step 4 — Scenario assessment:
The base case and favorable scenario (2 of 3) support Persistence.
The adverse scenario requires a specific external trigger not present in current data.
Without new information, the base case is the most probable outcome.

Step 5 — Judgment: Base case favors Persistence → (B).

</think>
<answer>(B)</answer>

Figure D.5: (Single-Prediction) Example of Event Response task.



Task: Support/Resistance — Ticker: MET — GT: (B) Bounce

Input: You are analyzing the stock MET. Below are the daily closing prices (120 days): [*prices shown above*]. The stock is currently near a key support level. Predict the behavior over the next 10 days.
(A) Breakout
(B) Bounce

Output:

<think>

Step 1 — Key level: support at 64.47.

Step 2 — Proximity: 1.9%.

Step 3 — Scenario analysis:
• *Base case* (most probable — no major external events):
The key level holds as institutional orders defend it. Price reverses direction after testing the level.

• *Adverse scenario* (external shock reverses expectation):
Sustained pressure eventually overwhelms the level, leading to a delayed breakout.

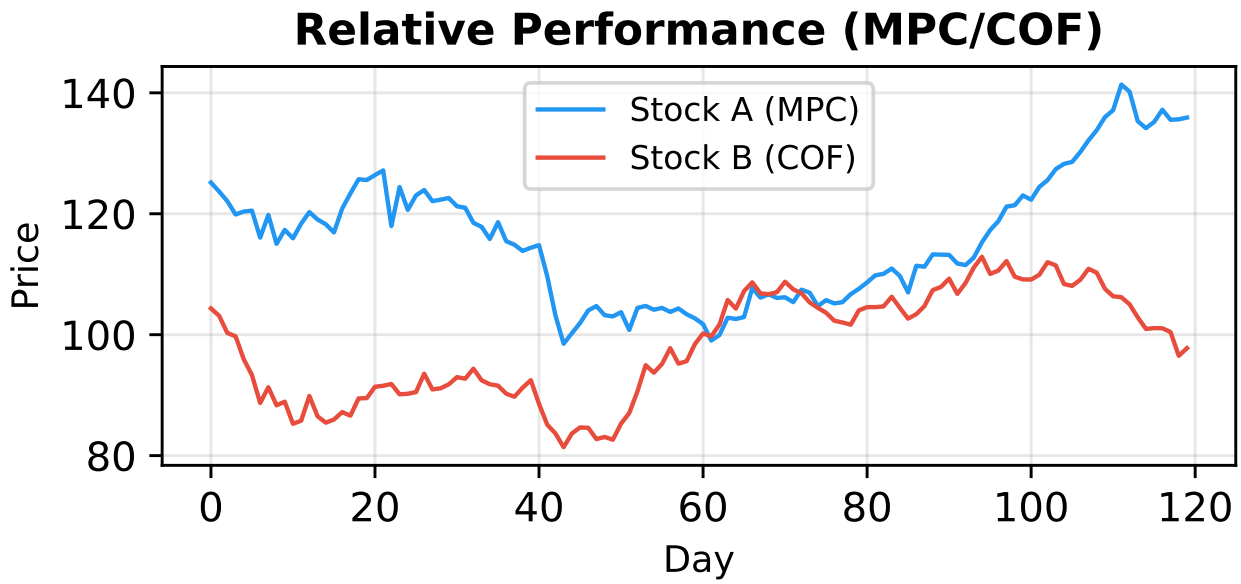
• *Favorable scenario* (catalyst strengthens expectation):
A strong bounce off the level triggers a significant reversal move in the opposite direction.

Step 4 — Scenario assessment:
The base case and favorable scenario (2 of 3) support Bounce.
The adverse scenario requires a specific external trigger not present in current data.
Without new information, the base case is the most probable outcome.

Step 5 — Judgment: Base case favors Bounce → (B).

</think>
<answer>(B)</answer>

Figure D.6: (Single-Prediction) Example of Support/Resistance task.



Task: Relative Performance — Tickers: MPC vs COF — GT: (A) Stock A outperforms

Input: You are comparing two stocks. Stock A (MPC) and Stock B (COF) daily closing prices (120 days each): [*prices shown above*].
Predict which stock will have a higher return over the next 20 days.
(A) Stock A outperforms
(B) Stock B outperforms

Output:

<think>

Step 1 — Comparing MPC vs COF.

Step 2 — Forward window: 20 days.

Step 3 — Scenario analysis:
• *Base case* (most probable — no major external events):
Stock A's recent strengthening continues, while Stock B's momentum fades. The relative performance gap widens further in A's favor.

• *Adverse scenario* (external shock reverses expectation):
A company-specific headwind hits Stock A (e.g., earnings miss, regulatory issue), eroding its lead.

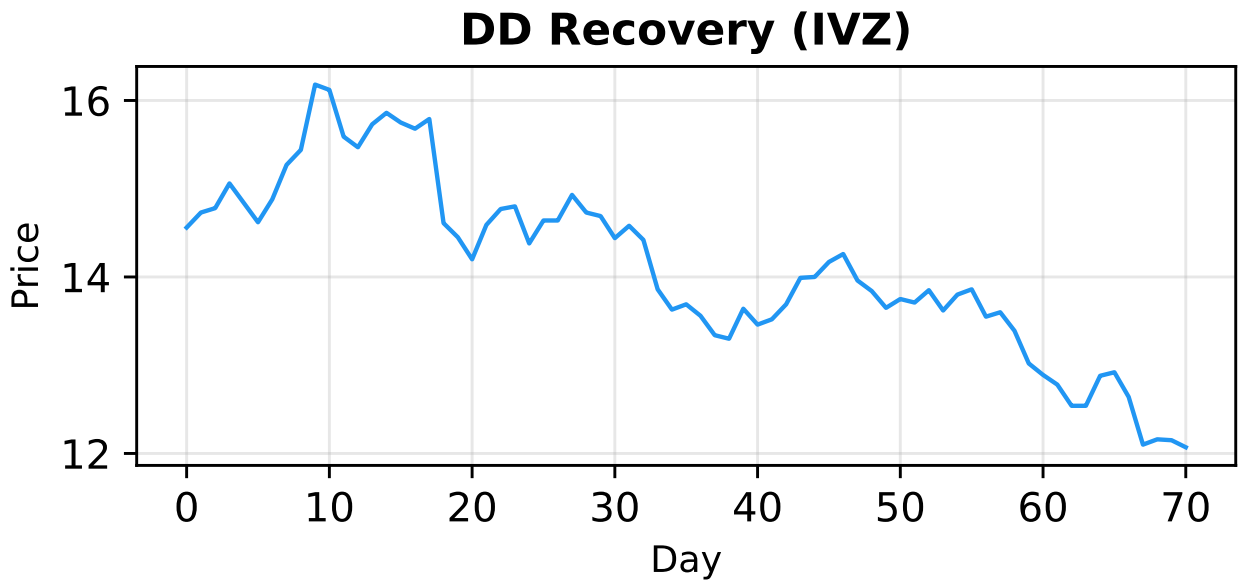
• *Favorable scenario* (catalyst strengthens expectation):
Sector tailwinds (e.g., favorable commodity prices) boost Stock A and widen its outperformance.

Step 4 — Scenario assessment:
The base case and favorable scenario (2 of 3) support Stock A outperforms.
The adverse scenario requires a specific external trigger not present in current data.
Without new information, the base case is the most probable outcome.

Step 5 — Judgment: Base case favors Stock A outperforms → (A).

</think>
<answer>(A)</answer>

Figure D.7: (Multi-Prediction) Example of Relative Performance task.



Task: Drawdown Recovery — Ticker: IVZ — GT: (B) Deepens

Input: You are analyzing the stock IVZ. Below are the daily closing prices: [*prices shown above*]. The stock has experienced a drawdown of 25.4% from its recent peak. Predict the behavior over the next 20 days.
(A) Recovery
(B) Deepens

Output:

<think>

Step 1 — Drawdown: 25.4% (peak=16.18, current=12.07).

Step 2 — Peak was 62 days ago.

Step 3 — Scenario analysis:
• *Base case* (most probable — no major external events):
Selling pressure continues as the fundamental weakness persists. The drawdown deepens before eventually finding a floor.

• *Adverse scenario* (external shock reverses expectation):
A surprise positive development reverses the decline — but this requires a specific external trigger.

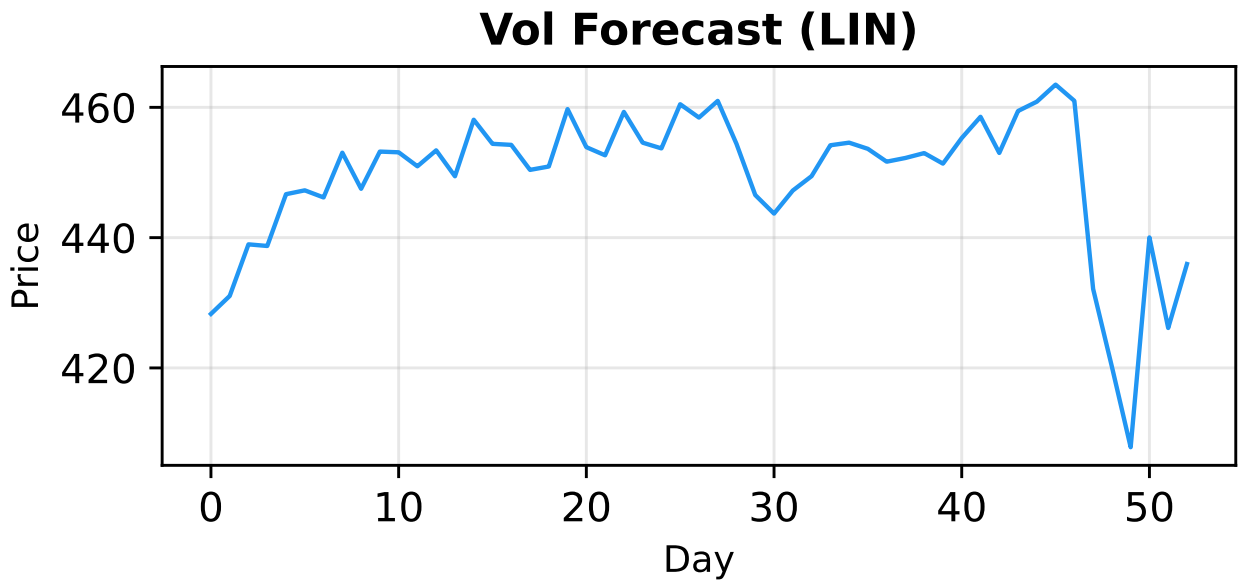
• *Favorable scenario* (catalyst strengthens expectation):
The decline accelerates as market-wide risk-off sentiment compounds the stock-specific weakness.

Step 4 — Scenario assessment:
The base case and favorable scenario (2 of 3) support Deepens.
The adverse scenario requires a specific external trigger not present in current data.
Without new information, the base case is the most probable outcome.

Step 5 — Judgment: Base case favors Deepens → (B).

</think>
<answer>(B)</answer>

Figure D.8: (Single-Prediction) Example of Drawdown Recovery task.



Task: Volatility Forecast — Ticker: LIN — GT: (B) Volatility decreases

Input: You are analyzing the stock LIN. Below are the daily closing prices: [*prices shown above*]. The current volatility ratio (recent 20-day vs overall) is 1.43. Predict how volatility will change over the next 20 days.
(A) Increases
(B) Decreases

Output:

<think>

Step 1 — Vol: recent=2.59%, long=1.81%, ratio=1.43.

Step 2 — Volatility pattern assessed.

Step 3 — Scenario analysis:
• *Base case* (most probable — no major external events):
Volatility mean-reverts toward the long-run level as the recent shock dissipates and the market normalizes.

• *Adverse scenario* (external shock reverses expectation):
A new source of uncertainty (earnings, macro news) triggers another volatility spike.

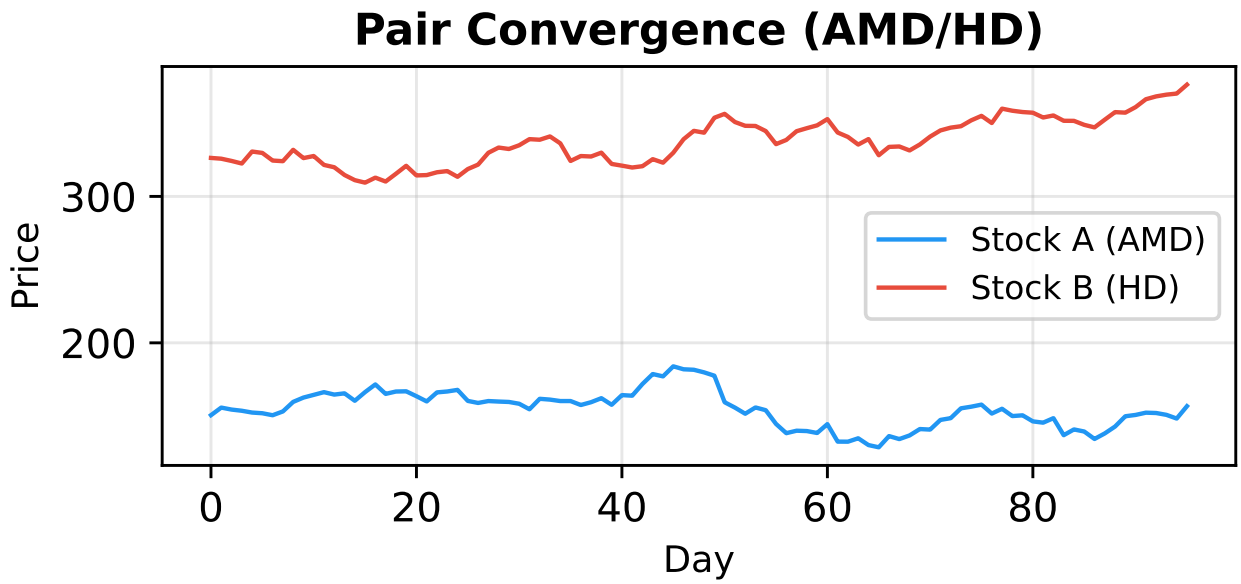
• *Favorable scenario* (catalyst strengthens expectation):
Market calming accelerates as uncertainty resolves, pushing volatility down faster than baseline.

Step 4 — Scenario assessment:
The base case and favorable scenario (2 of 3) support Vol decreases.
The adverse scenario requires a specific external trigger not present in current data.
Without new information, the base case is the most probable outcome.

Step 5 — Judgment: Base case favors Vol decreases → (B).

</think>
<answer>(B)</answer>

Figure D.9: (Single-Prediction) Example of Volatility Forecast task.



Task: Pair Convergence — Tickers: AMD vs HD — GT: (B) Divergence

Input: You are comparing two stocks. Stock A (AMD) and Stock B (HD) daily closing prices: [*prices shown above*]. The price spread (normalized difference) is 0.113. Predict how the spread will change over the next 20 days.
(A) Convergence
(B) Divergence

Output:

<think>

Step 1 — Pair: AMD, HD.

Step 2 — Current spread: 0.113.

Step 3 — Scenario analysis:
• *Base case* (most probable — no major external events):
Different fundamentals sustain the widening spread. The divergence continues as the two stocks respond to distinct sector dynamics.

• *Adverse scenario* (external shock reverses expectation):
A common catalyst (e.g., index rebalancing, macro shock) narrows the spread by forcing temporary co-movement.

• *Favorable scenario* (catalyst strengthens expectation):
Company-specific events accelerate the divergence further as one stock's fundamentals diverge from the other's.

Step 4 — Scenario assessment:
The base case and favorable scenario (2 of 3) support Divergence.
The adverse scenario requires a specific external trigger not present in current data.
Without new information, the base case is the most probable outcome.

Step 5 — Judgment: Base case favors Divergence → (B).

</think>
<answer>(B)</answer>

Figure D.10: (Multi-Prediction) Example of Pair Convergence task.

E Comparison with Other Methods

We present three qualitative examples per task where only FinSTaR answers correctly while both baselines (Qwen2.5-7B and TimeOmni-1-7B, zero-shot) fail.

E.1 Drawdown (Assessment)

Example 1: Drawdown (Assessment) — Ticker: ABBV — GT: (C)

Qwen2.5-7B (B) ✗

[think]
To determine the current drawdown phase of ABBV, we need to identify the recent peak price and then calculate how far the current price is from that peak.
1. Identify the highest price in the given data.
2. Calculate the difference between the highest price and the current price.
3. Determine the percentage decline based on the highest price.
From the data provided, the highest price is 17 [...]

TimeOmni-1 (B) ✗

[think]
Step 1: Identify the recent peak
- The data covers 120 trading days.
- The highest value in the series is 170.00 (last value), which is at index 119.
- The previous day's price is 168.75 (index 118).
- The day before that is 168.68 (index 117).
- The day before that is 165.54 (index 116).
- The day before that is 156.74 (index 115).
- The day before that is 158.70 (index 114) [...]

FinSTaR (Ours) (C) ✓

[think]
Step 1 – Find the peak price:
Scanning the 120-day series, the highest price is 170.00 (around day 108).
Step 2 – Current price:
The last price in the series is 152.33.
Step 3 – Calculate drawdown:
Drawdown = (Peak - Current) / Peak = (170.00 - 152.33) / 170.00 = 10.4 pct
Step 4 – Classify:
10.4 pct is between 10 pct and 20 pct, which corresponds to (C) Correction.
[/think]
[answer]C[/answer]

Example 2: Drawdown (Assessment) — Ticker: DHR — GT: (C)

Qwen2.5-7B (A) ✗

[think]
To determine the current drawdown phase of the stock DHR, we need to identify its recent peak price and then compare the current price to this peak.
1. Identify the Recent Peak Price:
- The data provided shows the most recent 120 trading days' closing prices.
- The highest value in this list is 211.53, which occurred on the last day of the provided data.
2. Determine the Curren [...]

TimeOmni-1 (A) ✗

[think]
Step 1: Identify the most recent peak
- The data shows a clear upward trend with some fluctuations, but the highest value in the series is at the end: 211.53 (last value).
- The previous high before this was 211.47 (second-to-last value), so the most recent peak is at or very close to the last value.
Step 2: Estimate the current price
- The most recent value is 211.53.
- The next-to-la [...]

FinSTaR (Ours) (C) ✓

[think]
Step 1 – Find the peak price:
Scanning the 120-day series, the highest price is 211.53 (around day 103).
Step 2 – Current price:
The last price in the series is 189.45.
Step 3 – Calculate drawdown:
Drawdown = (Peak - Current) / Peak = (211.53 - 189.45) / 211.53 = 10.5 pct
Step 4 – Classify:
10.5 pct is between 10 pct and 20 pct, which corresponds to (C) Correction.
[/think]
[answer]C[/answer]

Example 3: Drawdown (Assessment) — Ticker: CRWD — GT: (C)

Qwen2.5-7B (B) ✗

[think]
To determine the current drawdown phase of the stock CRWD, we need to identify the recent peak price and then compare the current price to this peak.
1. Identify the Recent Peak Price:
- The data provided shows the most recent 120 trading days' closing prices.
- The highest value in the list is 334.55, which occurred on one of the last few days.
2. Determine the Current Price:* [...]

TimeOmni-1 (B) ✗

[think]
Step 1: Identify the recent peak
- The data shows a clear upward trend with a few dips and rises.
- The highest value in the series is 334.55 (at index 95).
- The last value is 310.77 (index 119), which is lower than the peak but still relatively high.
Step 2: Estimate the current price
- The last value is 310.77.
- The previous value was 329.57 (index 118), so there has been a d [...]

FinSTaR (Ours) (C) ✓

[think]
Step 1 – Find the peak price:
Scanning the 120-day series, the highest price is 334.55 (around day 76).
Step 2 – Current price:
The last price in the series is 299.15.
Step 3 – Calculate drawdown:
Drawdown = (Peak - Current) / Peak = (334.55 - 299.15) / 334.55 = 10.6 pct
Step 4 – Classify:
10.6 pct is between 10 pct and 20 pct, which corresponds to (C) Correction.
[/think]
[answer]C[/answer]

E.2 Correlation (Assessment)

Example 1: Correlation (Assessment) — Ticker: AMD_BAC — GT: (A)

Qwen2.5-7B (C) ✗

[think]To determine the correlation between the daily returns of Stock A and Stock B, we need to calculate the correlation coefficient. However, given the complexity and the fact that we are asked to choose from predefined options, we can analyze the general trend visually or through a simplified approach.
Firstly, let's observe the trends:
- Stock A's prices show a significant fluctuation with so [...]

TimeOmni-1 (B) ✗

[think]
Step 1: Trend Consistency (Perceptual Matching)
- Both Stock A (AMD) and Stock B (BAC) start with relatively stable/flat periods at the beginning.
- Both show a gradual rise in the first half of the series (e.g., both increase from 137-144 for AMD and 44-46 for BAC).
- Both then experience a decline or stabilization period (e.g., AMD falls from 144 down to 110, BAC falls from 46 down [...]

FinSTaR (Ours) (A) ✓

[think]
Step 1 – Compare price changes in 12 periods (10 days each):
Days 1-10: A=+2.8 pct, B=+1.2 pct
Days 11-20: A=-14.3 pct, B=-7.2 pct
Days 21-30: A=+8.6 pct, B=+2.8 pct
Days 31-40: A=-2.8 pct, B=-0.6 pct
Days 41-50: A=-8.9 pct, B=+1.5 pct
Days 51-60: A=+3.6 pct, B=-5.1 pct
Days 61-70: A=-10.8 pct, B=-7.5 pct
Days 71-80: A=+10.0 pct, B=+2.6 pct
Days 81-90: A=-23.9 pct, B=-12.4 pct
Days 91-100: A=+3.1 pct, B=+8.8 pct
Days 101-110: A=+10.1 pct, B=+11.2 pct
Days 111-120: A=+20.0 pct, B=+8.7 pct
Step 2 – Pattern:
When A rises, B also tends to rise, and when A falls, B also falls. The magnitudes are similar.
Step 3 – positive correlation -> (A).
[/think]
[answer]A[/answer]

Example 2: Correlation (Assessment) — Ticker: CRM_DECK — GT: (A)

Qwen2.5-7B (B) ✗

[think]To determine the correlation between the daily returns of Stock A and Stock B, we need to calculate the correlation coefficient. However, given the complexity and the fact that we are only asked to determine the nature of the correlation (positive, negative, or no significant), we can infer this by observing the general trend of the price movements.
For Stock A (CRM):
- The prices show some [...]

TimeOmni-1 (B) ✗

[think]
Step 1: Trend Consistency (Perceptual Matching)
- Both Stock A (CRM) and Stock B (DECK) start with relatively stable values for the first few days.
- Around day 10-15, both begin to rise: CRM goes from 282 to 339, DECK from 152 to 178.
- Both peak around day 15-16, then both decline slightly.
- Both show a pronounced rise again starting around day 25-30: CRM rises from 320 to 365, DE [...]

FinSTaR (Ours) (A) ✓

[think]
Step 1 – Compare price changes in 12 periods (10 days each):
Days 1-10: A=+3.4 pct, B=+5.3 pct
Days 11-20: A=+8.3 pct, B=+7.8 pct
Days 21-30: A=+2.5 pct, B=+15.1 pct
Days 31-40: A=-4.6 pct, B=+3.0 pct
Days 41-50: A=-2.0 pct, B=+1.5 pct
Days 51-60: A=-2.5 pct, B=+1.3 pct
Days 61-70: A=+3.8 pct, B=-20.3 pct
Days 71-80: A=-4.4 pct, B=-8.1 pct
Days 81-90: A=-11.5 pct, B=-12.3 pct
Days 91-100: A=-4.5 pct, B=-15.3 pct
Days 101-110: A=-3.3 pct, B=-4.8 pct
Days 111-120: A=-6.0 pct, B=-11.8 pct
Step 2 – Pattern:
When A rises, B also tends to rise, and when A falls, B also falls. The magnitudes are similar.
Step 3 – positive correlation -> (A).
[/think]
[answer]A[/answer]

Example 3: Correlation (Assessment) — Ticker: MRK_EL — GT: (A)

Qwen2.5-7B (B) ✗

[think]To determine the correlation between the daily returns of Stock A and Stock B, we need to calculate the correlation coefficient. However, given the complexity and the fact that we are asked to choose from predefined options, we can analyze the general trend visually or through a simplified approach.
Firstly, let's observe the trends:
- Stock A's prices show fluctuations but generally seem t [...]

TimeOmni-1 (B) ✗

[think]
Step 1: Trend Consistency (Perceptual Matching)
- Both Stock A and Stock B start with relatively stable values for about the first 10-15 days.
- Around day 15-20, both begin to rise: Stock A goes from 91 to 97, while Stock B rises from 62 to 72.
- Both then continue rising, peaking around day 25-30 (A: 98, B: 78).
- After this peak, both decline: Stock A falls from 98 down to 85 by [...]

FinSTaR (Ours) (A) ✓

[think]
Step 1 – Compare price changes in 12 periods (10 days each):
Days 1-10: A=-2.7 pct, B=-24.0 pct
Days 11-20: A=-3.0 pct, B=+4.6 pct
Days 21-30: A=+4.0 pct, B=+14.0 pct
Days 31-40: A=-4.8 pct, B=-9.0 pct
Days 41-50: A=+1.6 pct, B=-0.8 pct
Days 51-60: A=-3.2 pct, B=+8.8 pct
Days 61-70: A=-7.8 pct, B=-22.2 pct
Days 71-80: A=+2.5 pct, B=+10.4 pct
Days 81-90: A=+3.7 pct, B=-1.9 pct
Days 91-100: A=-1.4 pct, B=-7.1 pct
Days 101-110: A=-11.8 pct, B=-21.5 pct
Days 111-120: A=-4.1 pct, B=+2.7 pct
Step 2 – Pattern:
When A rises, B also tends to rise, and when A falls, B also falls. The magnitudes are similar.
Step 3 – positive correlation -> (A).
[/think]
[answer]A[/answer]

E.3 Volatility Regime (Assessment)

Example 1: Volatility Regime (Assessment) — Ticker: BKNG — GT: (C)

Qwen2.5-7B (B) ✗

[think]
To determine the volatility regimes, we need to calculate both the recent (last 20 days) and overall (120-day) volatility. Volatility can be measured using the standard deviation of the returns.
1. Calculate the daily returns for each day.
2. Compute the standard deviation of these returns for the last 20 days (recent volatility).
3. Compute the standard deviation of these returns for the [...]

TimeOmni-1 (A) ✗

[think]
Step 1: Calculate the overall (120-day) volatility
- The first 120 values are given.
- Compute the standard deviation (sigma) of these 120 values.
Step 2: Calculate the recent (last 20 days) volatility
- Take the last 20 values from the list.
- Compute the standard deviation (sigma recent) of these 20 values.
Step 3: Compute the ratio
- Ratio = sigma recent / sigma overall
Step 4: Classify the vol [...]

FinSTaR (Ours) (C) ✓

[think]
Step 1 – Compute daily returns from prices:
e.g. (3441.67->3435.57): -0.17 pct, (3435.57->3414.34): -0.61 pct, (3414.34->3431.52): +0.50 pct
...
Step 2 – Overall volatility (std of all returns): 1.60 pct
Step 3 – Recent 20-day volatility: 2.59 pct
Step 4 – Ratio = 2.59 / 1.60 = 1.62
Recent volatility is well above the long-term average – unusually turbulent.
-> (C) high volatility.
[/think]
[answer]C[/answer]

Example 2: Volatility Regime (Assessment) — Ticker: BKNG — GT: (C)

Qwen2.5-7B (B) ✗

[think]
To determine the volatility regime, we need to calculate both the recent (last 20 days) and overall (120-day) volatility and then compute their ratio.
1. Calculate the overall volatility (120-day):
- First, find the mean of the 120-day closing prices.
- Then, calculate the standard deviation of these prices.
2. Calculate the recent volatility (last 20 days):
- Find the mea [...]

TimeOmni-1 (A) ✗

[think]
Step 1: Calculate the overall (120-day) volatility
- The first 120 values are given.
- To estimate volatility, we can use the standard deviation of these values.
Let's compute the mean and standard deviation for the full 120-day series:
Mean (approximate calculation):
Sum = 3086.97 + ... + 3496.08 + 3558.52 + 3578.10 + 3640.90 + 3769.80 + 3689.26 + 3720.72 + 3678.88 + 3669.75 + 3681.88 [...]

FinSTaR (Ours) (C) ✓

[think]
Step 1 – Compute daily returns from prices:
e.g. (3086.97->3083.60): -0.11 pct, (3083.60->3046.54): -1.20 pct, (3046.54->3085.46): +1.28 pct
...
Step 2 – Overall volatility (std of all returns): 1.62 pct
Step 3 – Recent 20-day volatility: 2.67 pct
Step 4 – Ratio = 2.67 / 1.62 = 1.65
Recent volatility is well above the long-term average – unusually turbulent.
-> (C) high volatility.
[/think]
[answer]C[/answer]

Example 3: Volatility Regime (Assessment) — Ticker: GL — GT: (C)

Qwen2.5-7B (A) ✗

[think]
To determine the volatility regimes, we need to calculate both the recent (last 20 days) and overall (120-day) volatilities and then compute their ratio.
1. Calculate the overall volatility (120-day):
- First, find the mean of the 120-day closing prices.
- Then, calculate the standard deviation of these prices.
2. Calculate the recent volatility (last 20 days):
- Find the [...]

TimeOmni-1 (B) ✗

[think]
Step 1: Calculate the overall volatility (120-day standard deviation)
First, I'll compute the mean of the entire series.
Summing up the first few values:
- The first 20 values are mostly in the 114-122 range.
- The next 20 values are in the 117-124 range.
- The next 20 values are in the 120-126 range.
- The next 20 values are in the 123-126 range.
- The next 20 values are in the 124-126 ra [...]

FinSTaR (Ours) (C) ✓

[think]
Step 1 – Compute daily returns from prices:
e.g. (114.11->114.12): +0.01 pct, (114.12->113.63): -0.43 pct, (113.63->114.89): +1.10 pct
...
Step 2 – Overall volatility (std of all returns): 4.51 pct
Step 3 – Recent 20-day volatility: 10.75 pct
Step 4 – Ratio = 10.75 / 4.51 = 2.38
Recent volatility is well above the long-term average – unusually turbulent.
-> (C) high volatility.
[/think]
[answer]C[/answer]

E.4 Trend Direction (Assessment)

Example 1: Trend Direction (Assessment) — Ticker: BKNG — GT: (B)

Qwen2.5-7B (A) ✗

[think]
To determine the trend direction, we need to analyze the overall change in the stock price over the 120-day period.
1. Calculate the first and last prices:
- First price: 2864.00
- Last price: 3501.24
2. Calculate the percentage change:
[
(Percentage Change) = (((Last Price) - (First Price))((First Price))) 100
]
[
(Percentage Change) = ((3501.24 - 2864.00)(2864.00 [...]

TimeOmni-1 (C) ✗

[think]
Step 1: Trend Consistency (Perceptual Matching)
Let's visually scan the series for broad structural trends.
- The first 40 data points show a general rise: from around 2864 up to about 3183 by day 12.
- There is a noticeable drop after day 12, with prices falling back toward 2700s by day 50.
- After day 50, there is another rise, peaking at about 3500 by day 110.
- The last 10-15 points s [...]

FinSTaR (Ours) (B) ✓

[think]
Step 1 – Read start and end prices:
Start (day 1): 2864.00
End (day 120): 3425.24
Step 2 – Compute cumulative return:
Return = (3425.24 - 2864.00) / 2864.00 = +19.6 pct
Step 3 – Classify:
+19.6 pct -> (B) Mild Uptrend (+5 pct to +20 pct).
[/think]
[answer]B[/answer]

Example 2: Trend Direction (Assessment) — Ticker: NVR — GT: (B)

Qwen2.5-7B (A) ✗

[think]
To determine the trend direction, we need to analyze the overall change in the stock price over the 120-day period.
1. Calculate the first and last prices:
- First price: 5173.62
- Last price: 6466.00
2. Calculate the percentage change:
[
(Percentage Change) = ((Last Price) - (First Price))((First Price)) 100
]
[
(Percentage Change) = (6466.00 - 5173.62)(5173.62) 100 [...]

TimeOmni-1 (A) ✗

[think]
Step 1: Trend Consistency (Perceptual Matching)
First, I visually scan the series for overall trend direction. The data starts at 5174 and rises steadily through the first half, with some fluctuations but generally rising. Around day 60-70, there is a noticeable flattening or slight decline, then another rise toward the end. There are no clear, sustained declines; the largest drops are on [...]

FinSTaR (Ours) (B) ✓

[think]
Step 1 – Read start and end prices:
Start (day 1): 5173.62
End (day 120): 6078.96
Step 2 – Compute cumulative return:
Return = (6078.96 - 5173.62) / 5173.62 = +17.5 pct
Step 3 – Classify:
+17.5 pct -> (B) Mild Uptrend (+5 pct to +20 pct).
[/think]
[answer]B[/answer]

Example 3: Trend Direction (Assessment) — Ticker: BLK — GT: (D)

Qwen2.5-7B (E) ✗

[think]
To determine the trend direction, we need to analyze the overall movement of the stock price over the 120-day period.
1. Calculate the percentage change from the first day to the last day:
- First day price: 1021.89
- Last day price: 800.99
- Percentage change = ((Last day price - First day price) / First day price) * 100
- Percentage change = ((800.99 - 1021.89) / 1021.89) [...]

TimeOmni-1 (E) ✗

[think]
Step 1: Trend Consistency (Perceptual Matching)
Let's visually scan the series for broad structural trends.
- The first 40 data points show a general rise: from around 1020 down to about 990, then up to 1030-1040, then back down to 980-990, and so on.
- There is a noticeable "bump" or local peak around index 22-24 (values near 1030-1040), followed by a fall.
- After that, there's a gradua [...]

FinSTaR (Ours) (D) ✓

[think]
Step 1 – Read start and end prices:
Start (day 1): 1021.89
End (day 120): 896.49
Step 2 – Compute cumulative return:
Return = (896.49 - 1021.89) / 1021.89 = -12.3 pct
Step 3 – Classify:
-12.3 pct -> (D) Mild Downtrend (-5 pct to -20 pct).
[/think]
[answer]D[/answer]

E.5 Event Response (Prediction)

Example 1: Event Response (Prediction) — Ticker: AMAT — GT: (B)

Qwen2.5-7B (A) ✗

[think]
To determine whether the price is likely to mean-revert or persist, we need to analyze the recent trends and the nature of the shock.
1. Trend Analysis:
- The data shows a mix of increases and decreases over the past 120 days, with no clear long-term upward or downward trend.
- There are periods of volatility, but overall, the stock price has been relatively stable around a centr [...]

TimeOmni-1 (A) ✗

[think]
Step 1: Identify the baseline trend and recent pattern
- The series shows a general upward trend with some volatility.
- There is a noticeable rise around the first half (e.g., from 180 to 193), then a decline, then another rise, and then a more pronounced drop towards the end (140s).
- The last few days show a sharp drop (from 180 down to 125), suggesting a major event or rever [...]

FinSTaR (Ours) (B) ✓

[think]
Step 1 – Event: positive shock, z=4.52, return=+13.70 pct.
Step 2 – Pre-event trend: downward.
Step 3 – Scenario analysis:
Base case (most probable – no major external events):
The shock reflects genuine new information being priced in. The trend continues as the market adjusts to the new fundamental value.
Adverse scenario (external shock reverses expectation):
Mean-reversion forces kick in – the initial shock proves to be an overreaction and the price reverts.
Favorable scenario (catalyst strengthens expectation):
Additional confirming news amplifies the original move, with momentum traders piling in.
Step 4 – Scenario assessment:
The base case and favorable scenario (2 of 3) support Persistence.
The adverse scenario requires a specific external trigger not present in current data.
Without new information, the base case is the most probable outcome.
Step 5 – Judgment: Base case favors Persistence -> (B).
[/think]
[answer]B[/answer]

Example 2: Event Response (Prediction) — Ticker: DXCM — GT: (B)

Qwen2.5-7B (A) ✗

[think]
To determine whether the price is likely to mean-revert or persist, we need to analyze the recent trends and the nature of the shock.
1. Trend Analysis:
- The data shows a mix of increases and decreases, but there's no clear long-term upward or downward trend over the past 120 days. There are periods of increase followed by decrease, indicating some volatility but no strong directio [...]

TimeOmni-1 (A) ✗

[think]
Step 1: Identify the baseline trend and recent pattern
- The series shows a general upward trend with some volatility.
- There is a noticeable rise around the middle third of the series (e.g., from 75 to 88), then a slight fall, then another rise, and finally a sharp spike at the end (last value: 81.62).
- The last few values before the spike were in the 80s, so the spike is a signif [...]

FinSTaR (Ours) (B) ✓

[think]
Step 1 – Event: positive shock, z=5.03, return=+15.23 pct.
Step 2 – Pre-event trend: downward.
Step 3 – Scenario analysis:
Base case (most probable – no major external events):
The shock reflects genuine new information being priced in. The trend continues as the market adjusts to the new fundamental value.
Adverse scenario (external shock reverses expectation):
Mean-reversion forces kick in – the initial shock proves to be an overreaction and the price reverts.
Favorable scenario (catalyst strengthens expectation):
Additional confirming news amplifies the original move, with momentum traders piling in.
Step 4 – Scenario assessment:
The base case and favorable scenario (2 of 3) support Persistence.
The adverse scenario requires a specific external trigger not present in current data.
Without new information, the base case is the most probable outcome.
Step 5 – Judgment: Base case favors Persistence -> (B).
[/think]
[answer]B[/answer]

Example 3: Event Response (Prediction) — Ticker: NOW — GT: (B)

Qwen2.5-7B (A) ✗

[think]
To determine whether the price is likely to mean-revert or persist, we need to analyze the recent trends and the magnitude of the shock.
1. Trend Analysis:
- The data shows a mix of increases and decreases, but there's no clear long-term upward or downward trend over the past 120 days.
- The last few days before the shock show a slight downward trend, with prices around 160-170.
[...]

TimeOmni-1 (A) ✗

[think]
Step 1: Identify the baseline trend and recent pattern
- The series shows a clear upward trend for the first 80 days, with a few dips but generally rising prices.
- Around day 80-90, there is a noticeable flattening or slight decline, suggesting a possible top or reversal.
- The last 10-15 days show a sharp drop, then a partial recovery, but still below previous highs.
Step 2: Assess th [...]

FinSTaR (Ours) (B) ✓

[think]
Step 1 – Event: positive shock, z=4.03, return=+12.93 pct.
Step 2 – Pre-event trend: downward.
Step 3 – Scenario analysis:
Base case (most probable – no major external events):
The shock reflects genuine new information being priced in. The trend continues as the market adjusts to the new fundamental value.
Adverse scenario (external shock reverses expectation):
Mean-reversion forces kick in – the initial shock proves to be an overreaction and the price reverts.
Favorable scenario (catalyst strengthens expectation):
Additional confirming news amplifies the original move, with momentum traders piling in.
Step 4 – Scenario assessment:
The base case and favorable scenario (2 of 3) support Persistence.
The adverse scenario requires a specific external trigger not present in current data.
Without new information, the base case is the most probable outcome.
Step 5 – Judgment: Base case favors Persistence -> (B).
[/think]
[answer]B[/answer]

E.6 Support/Resistance (Prediction)

Example 1: Support/Resistance (Prediction) — Ticker: BLK — GT: (A)

Qwen2.5-7B (B) ✗

[think]
To predict the stock's behavior, we need to analyze the recent trends and the current position relative to the resistance level.
1. Identify the Resistance Level: The resistance level appears to be around 1163.39, which is the highest price in the given data.
2. Recent Trends:
- The stock has been fluctuating around this resistance level.
- There have been attempts to break a [...]

TimeOmni-1 (B) ✗

[think]
Step 1: Identify the current trend and key levels
- The data shows a clear upward trend with some volatility.
- There is a noticeable rise starting around day 40-45, peaking at 1087.49 on day 49, then a decline to 1030.46 by day 52, followed by a new rise to 1167.06 by day 100.
- The last few days (days 91-100) show a plateau or slight uptrend, but no sharp breakouts.
- The most [...]

FinSTaR (Ours) (A) ✓

[think]
Step 1 – Key level: resistance at 1094.78.
Step 2 – Proximity: 2.4 pct.
Step 3 – Scenario analysis:
Base case (most probable – no major external events):
The price breaks through the key level with sustained momentum, establishing a new trading range beyond the previous barrier.
Adverse scenario (external shock reverses expectation):
The breakout fails – a false breakout traps traders and the price reverses sharply back through the level.
Favorable scenario (catalyst strengthens expectation):
The breakout is accompanied by high volume, leading to an accelerated move as stops are triggered.
Step 4 – Scenario assessment:
The base case and favorable scenario (2 of 3) support Breakout.
The adverse scenario requires a specific external trigger not present in current data.
Without new information, the base case is the most probable outcome.
Step 5 – Judgment: Base case favors Breakout -> (A).
[/think]
[answer]A[/answer]

Example 2: Support/Resistance (Prediction) — Ticker: MSFT — GT: (A)

Qwen2.5-7B (B) ✗

[think]
To predict the stock's behavior, we need to analyze the recent trends and the current position relative to the resistance level.
1. Identify the Resistance Level: The resistance level is a key price level where the stock has previously faced selling pressure, preventing it from rising further. From the data, it appears to be around 520 - 530.
2. Recent Trends:
- The stock has be [...]

TimeOmni-1 (B) ✗

[think]
Step 1: Identify the current context and trend
- The data covers 120 trading days, with the last value being 515.52.
- There is a clear upward trend visible, with several periods of rising prices and some minor dips.
- The last few values (last 10-15) show a slight downtrend, but overall, the price is still higher than at the start of the series.
- A "key resistance level" is mentioned, [...]

FinSTaR (Ours) (A) ✓

[think]
Step 1 – Key level: resistance at 510.74.
Step 2 – Proximity: 0.9 pct.
Step 3 – Scenario analysis:
Base case (most probable – no major external events):
The price breaks through the key level with sustained momentum, establishing a new trading range beyond the previous barrier.
Adverse scenario (external shock reverses expectation):
The breakout fails – a false breakout traps traders and the price reverses sharply back through the level.
Favorable scenario (catalyst strengthens expectation):
The breakout is accompanied by high volume, leading to an accelerated move as stops are triggered.
Step 4 – Scenario assessment:
The base case and favorable scenario (2 of 3) support Breakout.
The adverse scenario requires a specific external trigger not present in current data.
Without new information, the base case is the most probable outcome.
Step 5 – Judgment: Base case favors Breakout -> (A).
[/think]
[answer]A[/answer]

Example 3: Support/Resistance (Prediction) — Ticker: HCA — GT: (A)

Qwen2.5-7B (B) ✗

[think]
To predict the stock's behavior, we need to analyze the recent trends and the current position relative to the resistance level.
1. Identify the Resistance Level: The resistance level is a key price level where the stock has previously faced selling pressure, preventing it from rising further. From the data, the highest price in the last 120 days is around 339.58. We can assume this as [...]

TimeOmni-1 (B) ✗

[think]
Step 1: Identify the current context and trend
- The data covers 120 trading days, with the last value being 339.58.
- There is a clear upward trend visible, with several periods of rising prices and some minor dips.
- The last few values (last 10-15) show a slight decline after a period of higher prices, but overall, the trend is upward.
- A "key resistance level" is mentioned, which is [...]

FinSTaR (Ours) (A) ✓

[think]
Step 1 – Key level: resistance at 328.67.
Step 2 – Proximity: 3.3 pct.
Step 3 – Scenario analysis:
Base case (most probable – no major external events):
The price breaks through the key level with sustained momentum, establishing a new trading range beyond the previous barrier.
Adverse scenario (external shock reverses expectation):
The breakout fails – a false breakout traps traders and the price reverses sharply back through the level.
Favorable scenario (catalyst strengthens expectation):
The breakout is accompanied by high volume, leading to an accelerated move as stops are triggered.
Step 4 – Scenario assessment:
The base case and favorable scenario (2 of 3) support Breakout.
The adverse scenario requires a specific external trigger not present in current data.
Without new information, the base case is the most probable outcome.
Step 5 – Judgment: Base case favors Breakout -> (A).
[/think]
[answer]A[/answer]

E.7 Relative Performance (Prediction)

Example 1: Relative Performance (Prediction) — Ticker: MSCI_NVDA — GT: (A)

Qwen2.5-7B (B) ✗

[think]
To determine which stock is likely to outperform over the next 20 trading days, we need to analyze recent trends and patterns in the closing prices.
For Stock A (MSCI):
- The closing prices show a significant drop from around 580 to 434.85 over the first half of the period.
- There's a recovery phase where prices rise to around 500 but then decline again.
- The last few days show a slight [...]

TimeOmni-1 (B) ✗

[think]
Step 1: Trend Consistency (Perceptual Matching)
- Both MSCI and NVDA show a general upward trend in the first half of the series.
- MSCI: Starts around 530, rises to 587 by day 10, then fluctuates with some declines but remains above 500 for most of the period.
- NVDA: Starts at 57, rises to 92 by day 10, then fluctuates with some declines but remains above 70 for most of the pe [...]

FinSTaR (Ours) (A) ✓

[think]
Step 1 – Comparing MSCI vs NVDA.
Step 2 – Forward window: 20 days.
Step 3 – Scenario analysis:
Base case (most probable – no major external events):
Current momentum trends persist. Stock A continues its recent trajectory relative to Stock B, as momentum typically sustains over 20-day windows.
Adverse scenario (external shock reverses expectation):
Stock A faces a company-specific headwind (earnings miss, product recall) while Stock B benefits from a positive catalyst, reversing the relative performance.
Favorable scenario (catalyst strengthens expectation):
Sector tailwinds or index inclusion further boosts Stock A's advantage over Stock B.
Step 4 – Scenario assessment:
The base case and favorable scenario (2 of 3) support Stock A outperforms.
The adverse scenario requires a specific external trigger not present in current data.
Without new information, the base case is the most probable outcome.
Step 5 – Judgment: Base case favors Stock A outperforms -> (A).
[/think]
[answer](A)[/answer]

Example 2: Relative Performance (Prediction) — Ticker: CINF_GRMN — GT: (A)

Qwen2.5-7B (B) ✗

[think]To determine which stock is likely to outperform the other over the next 20 trading days, we need to analyze recent trends and patterns in the closing prices of both stocks.
For Stock A (CINF), the closing prices show a mix of increases and decreases over the past 120 days. However, there seems to be a general downward trend towards the end of the period, with some fluctuations around this [...]

TimeOmni-1 (B) ✗

[think]
Step 1: Trend Consistency (Perceptual Matching)
- Both CINF and GRMN start with relatively stable/flat periods for about the first 10-15 days.
- Around day 10-15, both begin to rise: CINF rises from 97 to 106, while GRMN rises from 90 to 96-97.
- Both then continue to rise, peaking around day 15-20: CINF peaks at 118, GRMN peaks at 98-99.
- After their respective peaks, both decline [...]

FinSTaR (Ours) (A) ✓

[think]
Step 1 – Comparing CINF vs GRMN.
Step 2 – Forward window: 20 days.
Step 3 – Scenario analysis:
Base case (most probable – no major external events):
Current momentum trends persist. Stock A continues its recent trajectory relative to Stock B, as momentum typically sustains over 20-day windows.
Adverse scenario (external shock reverses expectation):
Stock A faces a company-specific headwind (earnings miss, product recall) while Stock B benefits from a positive catalyst, reversing the relative performance.
Favorable scenario (catalyst strengthens expectation):
Sector tailwinds or index inclusion further boosts Stock A's advantage over Stock B.
Step 4 – Scenario assessment:
The base case and favorable scenario (2 of 3) support Stock A outperforms.
The adverse scenario requires a specific external trigger not present in current data.
Without new information, the base case is the most probable outcome.
Step 5 – Judgment: Base case favors Stock A outperforms -> (A).
[/think]
[answer](A)[/answer]

Example 3: Relative Performance (Prediction) — Ticker: NTRS_CTRA — GT: (A)

Qwen2.5-7B (B) ✗

[think]
To determine which stock is likely to outperform over the next 20 trading days, we need to analyze the recent trends and volatility of both stocks.
For Stock A (NTRS):
- The closing prices show a general upward trend with some fluctuations.
- There are periods of decline followed by recovery, indicating some volatility but overall positive movement.
For Stock B (CTRA):
- The closing prices [...]

TimeOmni-1 (B) ✗

[think]
Step 1: Trend Consistency (Perceptual Matching)
- Both NTRS and CTRA show a general downward trend in the first 20 days, then a stabilization or slight uptick around day 20-30, followed by a more pronounced rise in both stocks.
- Both stocks have a clear "stable -> rise" pattern after about day 30.
- The shapes match: both decline early, stabilize, then rise together.
- No major periods wh [...]

FinSTaR (Ours) (A) ✓

[think]
Step 1 – Comparing NTRS vs CTRA.
Step 2 – Forward window: 20 days.
Step 3 – Scenario analysis:
Base case (most probable – no major external events):
Current momentum trends persist. Stock A continues its recent trajectory relative to Stock B, as momentum typically sustains over 20-day windows.
Adverse scenario (external shock reverses expectation):
Stock A faces a company-specific headwind (earnings miss, product recall) while Stock B benefits from a positive catalyst, reversing the relative performance.
Favorable scenario (catalyst strengthens expectation):
Sector tailwinds or index inclusion further boosts Stock A's advantage over Stock B.
Step 4 – Scenario assessment:
The base case and favorable scenario (2 of 3) support Stock A outperforms.
The adverse scenario requires a specific external trigger not present in current data.
Without new information, the base case is the most probable outcome.
Step 5 – Judgment: Base case favors Stock A outperforms -> (A).
[/think]
[answer](A)[/answer]

E.8 Drawdown Recovery (Prediction)

Example 1: Drawdown Recovery (Prediction) — Ticker: AZO — GT: (A)

Qwen2.5-7B (B) ✗

[think]
To predict the stock's behavior over the next 20 trading days, we need to analyze the recent trends and patterns in the given data.
1. Identify the Recent Peak and Current Price: The recent peak is 3880.15, and the current price is 3603.17, indicating a significant drawdown of 7.1 pct.
2. Analyze Recent Trends:
- The last few prices show a fluctuating trend but generally moving [...]

TimeOmni-1 (B) ✗

[think]
Step 1: Identify the recent trend and key patterns
- The series shows a clear upward trend with some volatility, peaking at 3880.15 on day 89.
- There is a noticeable drop after this peak, reaching a trough at 3603.17 on day 90.
- The last 20 days show a gradual decline from the peak, with some minor fluctuations but no sharp upward moves.
Step 2: Estimate the direction of movement
- T [...]

FinSTaR (Ours) (A) ✓

[think]
Step 1 – Drawdown: 7.1 pct (peak=3880.15, current=3603.17).
Step 2 – Peak was 20 days ago.
Step 3 – Scenario analysis:
Base case (most probable – no major external events):
The drawdown stabilizes and the stock gradually recovers as value buyers accumulate shares. This is the typical pattern for moderate drawdowns in quality stocks.
Adverse scenario (external shock reverses expectation):
The underlying cause of the drawdown worsens (earnings revision, guidance cut), leading to further decline despite technical stabilization signals.
Favorable scenario (catalyst strengthens expectation):
A positive catalyst (analyst upgrade, strong earnings) triggers a sharp V-shaped recovery back to the peak.
Step 4 – Scenario assessment:
The base case and favorable scenario (2 of 3) support Recovery.
The adverse scenario requires a specific external trigger not present in current data.
Without new information, the base case is the most probable outcome.
Step 5 – Judgment: Base case favors Recovery -> (A).
[/think]
[answer]A[/answer]

Example 2: Drawdown Recovery (Prediction) — Ticker: NVR — GT: (A)

Qwen2.5-7B (B) ✗

[think]
To predict the stock's behavior, we need to analyze the recent trends and patterns in the given data.
1. Identify the Peak and Current Price: The peak price is 5967.22, and the current price is 5621.21, indicating a drawdown of 5.8 pct.
2. Analyze Recent Trends:
- The last few prices show a fluctuating trend but generally decreasing from around 5900 to 5600.
- There have been [...]

TimeOmni-1 (B) ✗

[think]
Step 1: Identify the recent trend and key patterns
- The series shows a clear upward trend with some volatility, peaking at 5967.22 on day 49.
- After the peak, there is a decline, reaching a local minimum around 5585.70 by day 66.
- The last 20 days show a gradual decline from 5887 to 5611, with no sharp upward moves.
Step 2: Estimate the current context
- The most recent value is 5 [...]

FinSTaR (Ours) (A) ✓

[think]
Step 1 – Drawdown: 5.8 pct (peak=5967.22, current=5621.21).
Step 2 – Peak was 26 days ago.
Step 3 – Scenario analysis:
Base case (most probable – no major external events):
The drawdown stabilizes and the stock gradually recovers as value buyers accumulate shares. This is the typical pattern for moderate drawdowns in quality stocks.
Adverse scenario (external shock reverses expectation):
The underlying cause of the drawdown worsens (earnings revision, guidance cut), leading to further decline despite technical stabilization signals.
Favorable scenario (catalyst strengthens expectation):
A positive catalyst (analyst upgrade, strong earnings) triggers a sharp V-shaped recovery back to the peak.
Step 4 – Scenario assessment:
The base case and favorable scenario (2 of 3) support Recovery.
The adverse scenario requires a specific external trigger not present in current data.
Without new information, the base case is the most probable outcome.
Step 5 – Judgment: Base case favors Recovery -> (A).
[/think]
[answer]A[/answer]

Example 3: Drawdown Recovery (Prediction) — Ticker: NVDA — GT: (A)

Qwen2.5-7B (B) ✗

[think]
To predict the stock's behavior, we need to analyze the recent trends and patterns in the given data.
1. Identify the Peak and Current Price: The peak price is 135.51, and the current price is 120.82, indicating a 10.8 pct drawdown from the peak.
2. Trend Analysis:
- The stock price has been fluctuating but generally shows a downward trend since reaching the peak.
- There hav [...]

TimeOmni-1 (B) ✗

[think]
Step 1: Identify the recent trend and key patterns
- The series shows a clear upward trend with several peaks and troughs.
- There is a notable rise starting around day 10-15, peaking at 135.51 (day 30), then a decline to 120.82 (day 98).
- The last 20 days show a gradual decline, with some minor fluctuations but no sharp upward moves.
Step 2: Estimate the current context
- The most [...]

FinSTaR (Ours) (A) ✓

[think]
Step 1 – Drawdown: 10.8 pct (peak=135.51, current=120.82).
Step 2 – Peak was 68 days ago.
Step 3 – Scenario analysis:
Base case (most probable – no major external events):
The drawdown stabilizes and the stock gradually recovers as value buyers accumulate shares. This is the typical pattern for moderate drawdowns in quality stocks.
Adverse scenario (external shock reverses expectation):
The underlying cause of the drawdown worsens (earnings revision, guidance cut), leading to further decline despite technical stabilization signals.
Favorable scenario (catalyst strengthens expectation):
A positive catalyst (analyst upgrade, strong earnings) triggers a sharp V-shaped recovery back to the peak.
Step 4 – Scenario assessment:
The base case and favorable scenario (2 of 3) support Recovery.
The adverse scenario requires a specific external trigger not present in current data.
Without new information, the base case is the most probable outcome.
Step 5 – Judgment: Base case favors Recovery -> (A).
[/think]
[answer]A[/answer]

E.9 Volatility Forecast (Prediction)

Example 1: Volatility Forecast (Prediction) — Ticker: BDX — GT: (A)

Qwen2.5-7B (B) ✗

[think]
To determine whether volatility will increase or decrease, we need to analyze the recent price movements and compare them with the historical volatility.
1. Recent Price Movements:
- The last 20 days' closing prices show a mix of increases and decreases, but there's no clear trend indicating a significant upward or downward movement.
- The recent volatility ratio is 0.91, which [...]

TimeOmni-1 (B) ✗

[think]
Step 1: Trend Consistency (Perceptual Matching)
- Looking at the series, there is a general downward trend with some fluctuations. The first half of the series shows a gradual decline from 185 to 170, then a slight uptick around day 20-25, followed by another decline and a small uptick near the end.
- The last 20 days (days 49-68) show a clear downtrend: prices start at 180 and graduall [...]

FinSTaR (Ours) (A) ✓

[think]
Step 1 – Vol: recent=1.06 pct, long=1.17 pct, ratio=0.91.
Step 2 – Volatility pattern assessed.
Step 3 – Scenario analysis:
Base case (most probable – no major external events):
Volatility clustering continues – the current elevated/rising volatility persists as uncertainty remains in the market.
Adverse scenario (external shock reverses expectation):
A resolution of the uncertainty driver (e.g., earnings release, policy decision) causes volatility to collapse rapidly.
Favorable scenario (catalyst strengthens expectation):
A new source of uncertainty emerges, compounding existing volatility and pushing it even higher.
Step 4 – Scenario assessment:
The base case and favorable scenario (2 of 3) support Vol increases.
The adverse scenario requires a specific external trigger not present in current data.
Without new information, the base case is the most probable outcome.
Step 5 – Judgment: Base case favors Vol increases -> (A).
[/think]
[answer](A)[/answer]

Example 2: Volatility Forecast (Prediction) — Ticker: KEY — GT: (A)

Qwen2.5-7B (B) ✗

[think]
To determine whether volatility will increase or decrease, we need to analyze the recent price movements and the given volatility ratio.
1. Price Trend Analysis:
- The closing prices show a mix of increases and decreases, but there isn't a clear upward or downward trend in the last 77 days.
- There are several periods where prices seem to fluctuate around a certain level before m [...]

TimeOmni-1 (B) ✗

[think]
Step 1: Trend Consistency (Perceptual Matching)
- Looking at the series, there is a clear downward trend in the first half (first 40 days), with prices gradually declining from around 10 to about 8. Then, after a brief flat period, there's a slight uptick in the last 37 days, but it's not as pronounced as the initial decline.
- The volatility ratio is given as 0.59, which means the recent [...]

FinSTaR (Ours) (A) ✓

[think]
Step 1 – Vol: recent=2.20 pct, long=3.74 pct, ratio=0.59.
Step 2 – Volatility pattern assessed.
Step 3 – Scenario analysis:
Base case (most probable – no major external events):
Volatility clustering continues – the current elevated/rising volatility persists as uncertainty remains in the market.
Adverse scenario (external shock reverses expectation):
A resolution of the uncertainty driver (e.g., earnings release, policy decision) causes volatility to collapse rapidly.
Favorable scenario (catalyst strengthens expectation):
A new source of uncertainty emerges, compounding existing volatility and pushing it even higher.
Step 4 – Scenario assessment:
The base case and favorable scenario (2 of 3) support Vol increases.
The adverse scenario requires a specific external trigger not present in current data.
Without new information, the base case is the most probable outcome.
Step 5 – Judgment: Base case favors Vol increases -> (A).
[/think]
[answer](A)[/answer]

Example 3: Volatility Forecast (Prediction) — Ticker: MRK — GT: (A)

Qwen2.5-7B (B) ✗

[think]
To determine whether volatility will increase or decrease, we need to analyze the recent price movements and compare them with the historical volatility.
1. Recent Price Movements: The last 20 days show a mix of increases and decreases, but generally, the prices have been trending upwards. The most recent prices are around 117-118, while the prices in the earlier part of the list were [...]

TimeOmni-1 (B) ✗

[think]
Step 1: Trend Consistency (Perceptual Matching)
- Looking at the series, there is a clear upward trend in the first half (days 0-49), with prices rising from 94 to 110, then a plateau, followed by a second rise (days 50-74), peaking around 120, then a gradual decline back toward 115, and finally a small uptick at the end.
- The last 20 days (days 75-94) show a gentle decline from 120 d [...]

FinSTaR (Ours) (A) ✓

[think]
Step 1 – Vol: recent=0.83 pct, long=1.02 pct, ratio=0.81.
Step 2 – Volatility pattern assessed.
Step 3 – Scenario analysis:
Base case (most probable – no major external events):
Volatility clustering continues – the current elevated/rising volatility persists as uncertainty remains in the market.
Adverse scenario (external shock reverses expectation):
A resolution of the uncertainty driver (e.g., earnings release, policy decision) causes volatility to collapse rapidly.
Favorable scenario (catalyst strengthens expectation):
A new source of uncertainty emerges, compounding existing volatility and pushing it even higher.
Step 4 – Scenario assessment:
The base case and favorable scenario (2 of 3) support Vol increases.
The adverse scenario requires a specific external trigger not present in current data.
Without new information, the base case is the most probable outcome.
Step 5 – Judgment: Base case favors Vol increases -> (A).
[/think]
[answer](A)[/answer]

E.10 Pair Convergence (Prediction)

Example 1: Pair Convergence (Prediction) — Ticker: PANW_CRWD — GT: (A)

Qwen2.5-7B (B) ✗

[think]
To determine whether the spread between PANW and CRWD will converge or diverge, we need to analyze the trends in both stock prices over the given period.
1. Trend Analysis for PANW:
- The closing prices of PANW show a general upward trend with some fluctuations.
- There are periods where the stock price increases significantly (e.g., from around 180 to 200+), followed by slight d [...]

TimeOmni-1 (B) ✗

[think]
Step 1: Trend Consistency (Perceptual Matching)
- Both PANW and CRWD show an initial rise in the first 10 days.
- Both then decline for about 10-15 days.
- Both then rise again, with CRWD showing a more pronounced increase after day 20.
- Both have periods of stability and minor fluctuations.
- The overall shapes match: rises, falls, and recoveries occur at similar times, though CRWD is a [...]

FinSTaR (Ours) (A) ✓

[think]
Step 1 – Pair: ['PANW', 'CRWD'].
Step 2 – Current spread: 0.325.
Step 3 – Scenario analysis:
Base case (most probable – no major external events):
The elevated spread narrows as the historical co-movement relationship reasserts itself. This is the typical mean-reversion pattern for correlated stocks.
Adverse scenario (external shock reverses expectation):
The divergence reflects a structural change (business pivot, sector reclassification) and the spread continues to widen.
Favorable scenario (catalyst strengthens expectation):
A common catalyst (sector earnings, index rebalancing) forces rapid convergence.
Step 4 – Scenario assessment:
The base case and favorable scenario (2 of 3) support Convergence.
The adverse scenario requires a specific external trigger not present in current data.
Without new information, the base case is the most probable outcome.
Step 5 – Judgment: Base case favors Convergence -> (A).
[/think]
[answer](A)[/answer]

Example 2: Pair Convergence (Prediction) — Ticker: ABBV_COIN — GT: (A)

Qwen2.5-7B (B) ✗

[think]
To determine whether the spread between ABBV and COIN will converge or diverge, we need to analyze the trends in their respective price movements.
1. Trend Analysis for ABBV:
- The closing prices of ABBV show a general upward trend with some fluctuations.
- There are periods where the stock price increases significantly (e.g., from 156.50 to 189.56), followed by periods of stabil [...]

TimeOmni-1 (B) ✗

[think]
Step 1: Trend Consistency (Perceptual Matching)
- Both ABBV and COIN show an initial decline in the first 10 days.
- Both then rise for about 10-15 days, peaking around day 25-30.
- Both then fall again, with a more pronounced drop for COIN starting around day 30-35.
- Both have a secondary rise after the mid-point, peaking again around day 50-55.
- Both then fall again, with a smaller di [...]

FinSTaR (Ours) (A) ✓

[think]
Step 1 – Pair: ['ABBV', 'COIN'].
Step 2 – Current spread: 0.346.
Step 3 – Scenario analysis:
Base case (most probable – no major external events):
The elevated spread narrows as the historical co-movement relationship reasserts itself. This is the typical mean-reversion pattern for correlated stocks.
Adverse scenario (external shock reverses expectation):
The divergence reflects a structural change (business pivot, sector reclassification) and the spread continues to widen.
Favorable scenario (catalyst strengthens expectation):
A common catalyst (sector earnings, index rebalancing) forces rapid convergence.
Step 4 – Scenario assessment:
The base case and favorable scenario (2 of 3) support Convergence.
The adverse scenario requires a specific external trigger not present in current data.
Without new information, the base case is the most probable outcome.
Step 5 – Judgment: Base case favors Convergence -> (A).
[/think]
[answer](A)[/answer]

Example 3: Pair Convergence (Prediction) — Ticker: ABBV_META — GT: (A)

Qwen2.5-7B (B) ✗

[think]
To determine whether the spread between ABBV and META will converge or diverge, we need to analyze the trends in their respective price movements.
1. Trend Analysis for ABBV:
- The closing prices of ABBV show a significant fluctuation but generally seem to be decreasing over time. There are periods of slight increases, but overall, the trend is downward.
2. Trend Analysis for META [...]

TimeOmni-1 (B) ✗

[think]
Step 1: Trend Consistency (Perceptual Matching)
- Both ABBV and META show an initial rise in the first 10-15 days.
- ABBV: Starts 144.80, rises to 149.32 by day 10, then falls back to 134.80 by day 15.
- META: Starts 212.18, rises to 241.29 by day 15, then falls back to 231.70 by day 20.
- Both then fall for about 10-15 days, with some minor fluctuations.
- ABBV: Drops fr [...]

FinSTaR (Ours) (A) ✓

[think]
Step 1 – Pair: ['ABBV', 'META'].
Step 2 – Current spread: 0.547.
Step 3 – Scenario analysis:
Base case (most probable – no major external events):
The elevated spread narrows as the historical co-movement relationship reasserts itself. This is the typical mean-reversion pattern for correlated stocks.
Adverse scenario (external shock reverses expectation):
The divergence reflects a structural change (business pivot, sector reclassification) and the spread continues to widen.
Favorable scenario (catalyst strengthens expectation):
A common catalyst (sector earnings, index rebalancing) forces rapid convergence.
Step 4 – Scenario assessment:
The base case and favorable scenario (2 of 3) support Convergence.
The adverse scenario requires a specific external trigger not present in current data.
Without new information, the base case is the most probable outcome.
Step 5 – Judgment: Base case favors Convergence -> (A).
[/think]
[answer](A)[/answer]