HINDSIGHT PREFERENCE OPTIMIZATION FOR FINANCIAL TIME SERIES ADVISORY

Yanwei Cui[1] **Guanghui Wang**[1] **Xing Zhang**[1] **Peiyang He**[1*]
Ziyuan Li[2] **Bing Zhu**[2] **Wei Qiu**[2] **Xusheng Wang**[2] **Zheng Yu**[2] **Anqi Xin**[2]

[1]AWS Generative AI Innovation Center [2]HSBC Holdings Plc. HSBC Technology Center China

ABSTRACT

Time series models predict numbers; decision-makers need advisory—directional signals with reasoning, actionable suggestions, and risk management. Training language models for such predictive advisory faces a fundamental challenge: quality depends on outcomes unknown at prediction time. We bridge two ideas from reinforcement learning—using information unavailable during execution to retrospectively generate training signal, and preference alignment—and propose Hindsight Preference Optimization: observed outcomes let an LLM judge rank candidate advisories on dimensions that scalar metrics cannot capture, producing preference pairs for DPO without human annotation. We apply this to Vision-Language-Model-based predictive advisories on S&P 500 equity time series, demonstrated by a 4B model outperforming its 235B teacher on both accuracy and advisory quality.

Track: Research

1 INTRODUCTION

Recent work has shown that language model architectures transfer remarkably well to time series forecasting (Ansari et al., 2024; Das et al., 2024; Shi et al., 2025), positioning pretrained models as powerful tools for temporal prediction. Yet these models output tokenized numbers, not natural language. Decision-makers need more: not just "the price will be $152" but why (reasoning about patterns), how confident (calibrated uncertainty), and what to do (actionable suggestions with risk management). This structured advisory requires natural language generation, motivating language models that can process historical data and produce reasoned, actionable output.

Vision-language models (VLMs) offer a natural bridge: they can interpret financial charts and generate textual analysis (Bai et al., 2025). However, training VLMs for predictive advisory introduces a challenge distinct from typical vision-language tasks. Unlike image captioning or visual question answering, where quality can be judged immediately, advisory quality depends on outcomes that unfold after prediction.

Our approach bridges two ideas from reinforcement learning: the retrospective training signal—using information unavailable during execution to generate supervision after the fact (Andrychowicz et al., 2017)—and preference alignment, where RLHF and its scalable variants—RLAIF (Lee et al., 2024), LLM-as-a-Judge (Zheng et al., 2023)—align models using preference pairs. We extend this framework by providing the judge with ground-truth outcomes, enabling objective evaluation of predictive advisory. The resulting framework, Hindsight Preference Optimization, automatically constructs preference pairs for DPO (Rafailov et al., 2023) from observed outcomes—eliminating human annotation while rewarding analytical quality over coincidental correctness.

Our contributions: (1) **Hindsight Preference Optimization**—grounding LLM-as-a-Judge evaluation in observed outcomes to assess structured predictions—scenarios, risk estimates, and reasoning—beyond what scalar metrics can capture; and (2) **VLM-based financial advisory**—applying our framework to train a 4B model that interprets candlestick charts and generates reasoned advisory, outperforming its 235B teacher on both accuracy and advisory quality.

*Corresponding author: peiyan@amazon.com

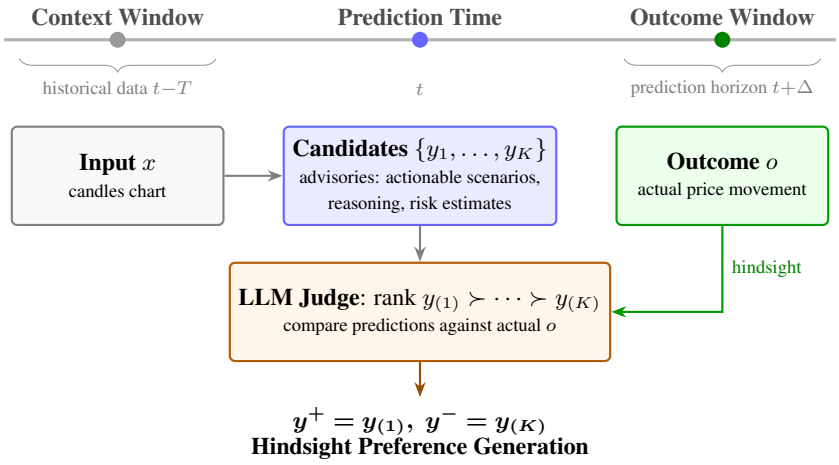

Figure 1: Hindsight Preference Optimization framework. The LLM judge ranks candidate advisories against observed outcomes, producing preference pairs for DPO.

2 Related Work

Time Series and Language Models. Recent work reveals that language model architectures generalize effectively to temporal data. Chronos (Ansari et al., 2024) and TimesFM (Das et al., 2024) demonstrate that tokenizing time series values enables transformers to achieve strong zero-shot forecasting. More relevant to our work, Kronos (Shi et al., 2025) targets financial markets specifically, pre-training on billions of candlestick records to learn price dynamics and trading patterns—though it outputs numerical predictions rather than natural language advisory. Vision-language models extend forecasting to visual inputs: Qwen3-VL (Bai et al., 2025) generates textual analysis from images. In finance, domain-specific LLMs have shown promise for market text analysis (Li et al., 2023). We build on these advances, combining chart understanding with structured advisory generation.

Preference Learning with Hindsight. Direct Preference Optimization (Rafailov et al., 2023) enables training from preference pairs without explicit reward modeling. RLAIF (Lee et al., 2024) and LLM-as-a-Judge (Zheng et al., 2023) scale preference collection through AI feedback, proving effective for subjective qualities. The idea of retrospective training signal—using post-hoc knowledge to generate supervision unavailable during execution—originates in reinforcement learning (Andrychowicz et al., 2017). Recent work such as mDPO (Wang et al., 2024) and RPO (Zhao & Li, 2025) confirms that enriching DPO pairs with external signal improves alignment in vision-language models for reducing hallucination. We bring this principle to preference learning, grounding evaluation in observed outcomes for predictive tasks.

3 Method

3.1 Problem Setup

Let $x \in \mathcal{X}$ denote the input context, which may include multiple modalities (e.g., time series data, charts, text). Let $y \in \mathcal{Y}$ denote the generated advisory containing predictive claims—directional forecasts, confidence levels, specific targets—along with reasoning and actionable guidance.

We also have $o \in \mathcal{O}$, the ground-truth outcome, which is unavailable at prediction time but can be collected after the prediction window. Unlike standard forecasting that minimizes prediction error against o, we use o to enable comparative evaluation: the judge ranks advisories by how well they anticipated the outcome, producing preference pairs for DPO rather than scalar losses.

3.2 HINDSIGHT PREFERENCE GENERATION

Given candidates $\{y_1, \ldots, y_K\}$ generated from input x (Figure 1), an LLM-as-a-Judge J evaluates each against the ground-truth outcome o, producing a ranking $y_{(1)} \succ \cdots \succ y_{(K)}$. Each advisory contains freeform reasoning alongside structured predictions: multiple scenarios, each describing a predicted price path (e.g., steady decline, V-shaped recovery) with an associated likelihood, as well as quantitative risk estimates such as expected drawdown and volatility. These are not assumed to be calibrated—they are the model's expressed assessments, which the judge evaluates against realized outcomes. Traditional scalar metrics such as MAE or MSE can evaluate point predictions but cannot assess reasoning quality, scenario calibration, or risk management; an LLM-as-a-Judge operating with hindsight can evaluate these richer dimensions jointly.

With access to outcomes unavailable at prediction time, the judge can distinguish sound analysis from lucky guesses—the core advantage of hindsight. It jointly evaluates *directional signals* (scenario-outcome alignment), *reasoning* (whether identified dynamics actually drove the move), and *risk management* (drawdown and volatility calibration). The resulting preference pairs encode not just *what* to predict but *how* to reason, derived automatically without human annotation.

3.3 HINDSIGHT-GUIDED PREFERENCE OPTIMIZATION

Stage 1: Supervised Fine-Tuning. We adopt vision-language models (VLMs) as the base architecture, leveraging their pretrained capabilities for chart understanding and textual reasoning. A large teacher model generates K candidate advisories for each input x. Using hindsight ranking, the judge selects top-ranked candidates as supervision targets y^*. A smaller student model is trained to reproduce these hindsight-selected examples:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(x,y^*)} \left[\log \pi_\theta(y^* \mid x)\right] \tag{1}$$

This bootstraps the student from the teacher's best candidates.

Stage 2: Hindsight DPO. We refine the student model through on-policy preference optimization: (1) sample K candidates from the student model, (2) rank against outcomes using the judge, (3) construct preference pairs with $y^+ = y_{(1)}$ and $y^- = y_{(K)}$, and (4) optimize using DPO. This process naturally extends to multiple rounds of iterative refinement:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(x,y^+,y^-)} \left[\log \sigma \left(\beta \left(\log \frac{\pi_\theta(y^+ \mid x)}{\pi_{\text{ref}}(y^+ \mid x)} - \log \frac{\pi_\theta(y^- \mid x)}{\pi_{\text{ref}}(y^- \mid x)}\right)\right)\right] \tag{2}$$

We augment with anchored regularization on chosen responses (Wang et al., 2024; Zhao & Li, 2025), following recent practices in preference optimization:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{DPO}}(\theta) + \alpha \cdot \mathcal{L}_{\text{SFT}}(y^+ \mid x) \tag{3}$$

The DPO term isolates what distinguishes y^+ from y^-—shared surface-level variations cancel out in the log-likelihood ratio, leaving the patterns that drive preference. The SFT anchor on y^+ prevents likelihood degradation as preferences shift the distribution.

4 EXPERIMENTS AND DISCUSSION

Setup. We evaluate our framework on S&P 500 equities using daily OHLC (Open-High-Low-Close) data spanning 2013–2017 (Nugent, 2018). Each sample pairs a candlestick chart showing one month of historical trading with the ground-truth price movements from the subsequent week. Charts contain only price candles and volume bars with no ticker symbols, dates, or axis labels that would identify the security or time period. The model generates structured advisory for this future window: directional scenarios with reasoning and risk estimates. We select the top 5 S&P 500 stocks and use a temporal split: data from 2013–2016 serves as the training set (1,395 samples at 5-day intervals), while all of 2017 is held out for testing (365 samples at 5-day intervals). We use Qwen3-VL-235B as the teacher model for generating candidate advisories during SFT data collection, and Qwen3-VL-4B as the student model that undergoes training. For preference pair construction (both SFT selection and DPO ranking), the LLM judge is Pixtral Large 25.02; for pairwise preference evaluation (Table 2), we use Claude Opus 4. Training uses LoRA on 4× L40S GPUs (g6e.12xlarge) with batch size 16 for both SFT and DPO.

Table 1: Prediction quality on held-out 2017 data (mean ± std, 5 evaluation runs).

Model	**Directional Acc.**	**Top-1 Scenario Acc.**
Claude Sonnet 3.7 (zero-shot)	56.1% ± 0.8%	27.3% ± 2.4%
Qwen3-VL-235B (zero-shot)	51.9% ± 1.4%	23.7% ± 1.5%
Qwen3-VL-4B (zero-shot)	50.7% ± 1.8%	22.1% ± 1.3%
Qwen3-VL-4B + SFT (Stage 1)	53.3% ± 2.7%	25.2% ± 1.8%
Qwen3-VL-4B + Hindsight DPO	57.9% ± 2.3%	27.1% ± 2.0%

Table 2: Pairwise preference win rates on held-out 2017 data (mean ± std, 5 evaluation runs).

	Win Rate Against		
Model	**Qwen3-VL-235B**	**Claude Sonnet 3.7**	**Qwen3-VL-4B**
Qwen3-VL-4B + SFT (Stage 1)	48.7% ± 1.3%	46.7% ± 1.4%	62.5% ± 1.8%
Qwen3-VL-4B + Hindsight DPO	56.8% ± 2.7%	52.6% ± 0.8%	71.4% ± 2.1%

We assess performance using three metrics. *Directional Accuracy*: percentage of correct directional predictions, excluding moves within ±1% as ambiguous (235 of 365 test samples). *Top-1 Scenario Accuracy*: whether the model's top predicted scenario (e.g., `RALLY_THEN_FADE`, `V_RECOVERY`, among 5 labels) matches the realized 5-day price path; computed on all 365 test samples. *Pairwise Preference*: an LLM judge with access to outcomes ranks advisory from two models by actionable guidance, reasoning, and risk management. We report means over 5 evaluation runs.

Results. Table 1 and Table 2 present our main results. Per-class precision/recall, dataset distribution, example model inputs, and a detailed advisory comparison are provided in Appendices A–C.

SFT on hindsight-selected examples teaches the 4B student structured reasoning and advisory format, achieving 53.3% directional accuracy and 25.2% scenario accuracy (Table 1)—both comparable to the 235B teacher (51.9% and 23.7%; McNemar's test, $p>0.05$). Pairwise preference confirms parity: the SFT model achieves only 48.7% win rate against the teacher (Table 2). Even boosted with top-ranked candidates, SFT distills the teacher's capabilities but cannot surpass it—the contrastive signal needed to distinguish better reasoning is absent.

Hindsight DPO bridges this gap. By contrasting advisories ranked against observed outcomes, DPO shifts the model toward patterns the judge associates with better predictive advisory. After preference optimization, the model achieves 57.9% directional accuracy and 27.1% scenario accuracy, significantly outperforming both SFT and the 235B teacher on both metrics (McNemar's test, $p<0.05$). Pairwise win rates also improve to 56.8% against Qwen3-VL-235B. As reference points, the improvement over the 4B zero-shot base is decisive (71.4% win rate), and the DPO-trained model achieves a 52.6% win rate against Claude Sonnet 3.7 (sign test, $p<0.05$).

Limitations. Our evaluation is limited to S&P 500 equities using only visual price and volume patterns; market outcomes also depend on news, earnings, and macroeconomic factors not captured in charts. The 2013–2017 evaluation window is predominantly bullish, and performance in volatile or bearish regimes (e.g., 2008, 2020, 2022) remains untested. Our goal is not to outperform the market but to improve the quality of structured advisory—reasoning, scenario calibration, and risk awareness—for decision support. Furthermore, both preference construction and evaluation rely on LLM judges without human expert validation; real-world financial adoption would require domain expert assessment to ensure advisory quality aligns with practitioner standards. Additionally, the current approach uses only the judge's rankings, which provides a weak learning signal for DPO—leading to slow and unstable convergence. Incorporating judge rationales as natural-language reward signals and disentangling which advisory dimensions—scenario calibration, reasoning quality, or risk estimation—drive improvement are promising directions.

5 Conclusion

Hindsight Preference Optimization bridges two ideas from reinforcement learning—retrospective training signal and preference alignment—for predictive advisory. Armed with observed outcomes, an LLM-as-a-Judge evaluates structured advisory along dimensions inaccessible to scalar metrics: directional signals, reasoning, actionable suggestions, and risk management. The resulting preference pairs for DPO require no human annotation. Our experiments show that prediction accuracy and advisory quality can be jointly optimized, enabling a 4B model to outperform its 235B teacher on both dimensions. The framework applies broadly: any predictive domain where outcomes are eventually observed can benefit from hindsight-grounded preference learning.

References

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. In *Advances in Neural Information Processing Systems*, 2017.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle Maddix Robinson, Hao Wang, Michael Mahoney, Kari Torkkola, Andrew Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, et al. Qwen3-vl technical report, 2025. URL `https://arxiv.org/abs/2511.21631`.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Forty-first International Conference on Machine Learning*, 2024.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. In *International Conference on Machine Learning*, pp. 26874–26901. PMLR, 2024.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. Large language models in finance: A survey. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pp. 374–382, 2023.

Cam Nugent. S&p 500 stock data. Kaggle, 2018. URL `https://www.kaggle.com/datasets/camnugent/sandp500`. CC0: Public Domain.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023.

Yu Shi, Zongliang Fu, Shuo Chen, Bohan Zhao, Wei Xu, Changshui Zhang, and Jian Li. Kronos: A foundation model for the language of financial markets. In *NeurIPS 2025 Workshop: Generative AI in Finance*, 2025.

Fei Wang, Wenxuan Zhou, James Y. Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mDPO: Conditional preference optimization for multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8078–8088, 2024.

Zihui Zhao and Zechang Li. Reflective preference optimization (rpo): Enhancing on-policy alignment via hint-guided reflection. *arXiv preprint arXiv:2512.13240*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

A Dataset Details

Table 3 summarizes the held-out 2017 evaluation set. The 365 samples across 5 tickers (AAPL, AMZN, FB, GOOGL, MSFT; 73 per ticker) exhibit a bullish skew reflective of the 2017 market: 43.8% of 5-day windows are classified as Bullish ($\geq$1% gain), 20.5% as Bearish ($\leq$-1% loss), and 35.6% as Neutral. After excluding the 130 neutral samples, the 235 directional samples used for accuracy evaluation have a 68.1%/31.9% Bullish/Bearish class imbalance.

Table 3: Evaluation set distribution (held-out 2017, 365 samples across 5 tickers: AAPL, AMZN, FB, GOOGL, MSFT).

Direction	**Criterion**	**Count**	**%**
Bullish	$\geq$ +1%	160	43.8%
Bearish	$\leq -1\%$	75	20.5%
Neutral (excluded)	within $\pm$1%	130	35.6%
Directional (evaluated)		**235**	**64.4%**

Table 4 reports per-class precision and recall on the 235 directional samples.

Table 4: Per-class precision and recall on held-out 2017 directional samples (mean $\pm$ std, 5 runs).

	Bullish		**Bearish**	
Model	**Prec.**	**Recall**	**Prec.**	**Recall**
Claude Sonnet 3.7 (zero-shot)	66.5 $\pm$ 0.7	71.5 $\pm$ 1.0	27.6 $\pm$ 2.2	23.2 $\pm$ 2.8
Qwen3-VL-235B (zero-shot)	66.2 $\pm$ 1.2	60.0 $\pm$ 1.2	28.9 $\pm$ 2.1	34.7 $\pm$ 3.3
Qwen3-VL-4B + SFT (Stage 1)	66.3 $\pm$ 1.8	63.8 $\pm$ 3.4	28.6 $\pm$ 3.4	30.9 $\pm$ 4.0
Qwen3-VL-4B + Hindsight DPO	**68.0** $\pm$ 1.5	**72.0** $\pm$ 3.0	**31.7** $\pm$ 3.9	27.7 $\pm$ 4.8

B Example Input

Figure 2 shows the candlestick chart provided to the VLM at inference time (20 trading days of historical data). The model generates structured advisory based solely on this visual input—no ticker symbols, dates, or axis labels that would identify the security or time period are provided.

Figure 3 shows the extended chart used by the LLM judge during training. The shaded region (Days 21–25) reveals the 5-day outcome, enabling the judge to evaluate advisory quality with hindsight and construct preference pairs for DPO.

C Advisory Comparison

We compare advisories generated by three models on the same input (the candlestick chart in Figure 2). The ground truth outcome was **BEARISH** (-2.72%, max drawdown -4.26%). Table 5 summarizes the predictions.

Table 5: Advisory comparison on a held-out AAPL sample. Ground truth: BEARISH (-2.72%, max drawdown -4.26%).

	Qwen3-VL-235B	**Claude Sonnet 3.7**	**Hindsight DPO (4B)**
Outlook	BULLISH	BULLISH	**BEARISH** ✓
Top scenario	STEADY_UP (55%)	RALLY_THEN_FADE (40%)	RALLY_THEN_FADE (45%)
Selloff probability	5%	10%	35%
Max drawdown est.	-2.5%	-3.5%	-2.5%

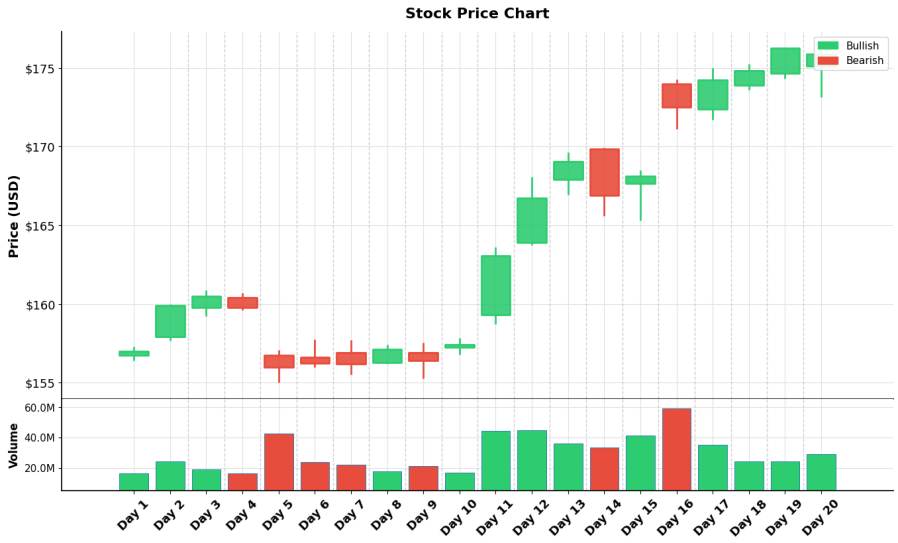

Figure 2: Model input: 20-day candlestick chart with price and volume. The model generates advisory from this chart alone.

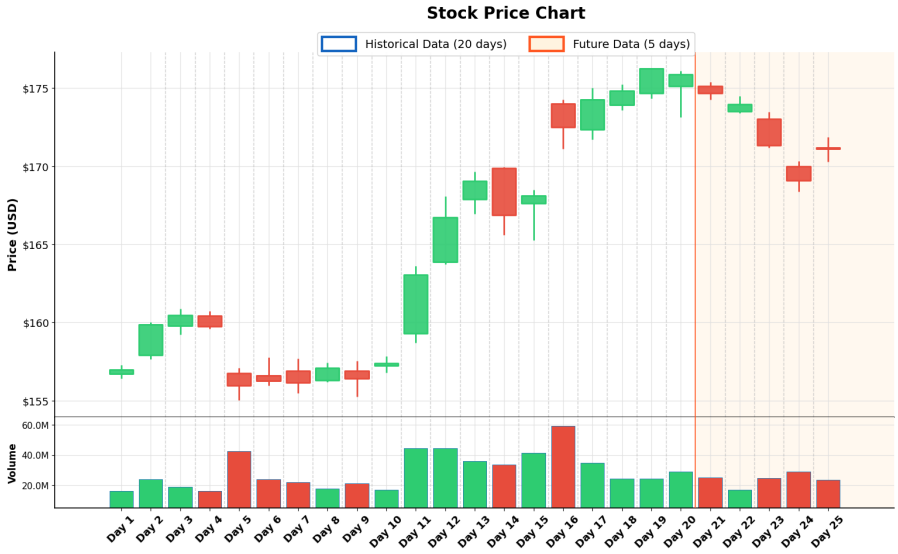

Figure 3: Judge input: the same chart extended with the 5-day outcome window (shaded region, Days 21–25). The judge uses this hindsight to rank candidate advisories.

Below we show the full advisory output from each model, including directional signals, reasoning, actionable suggestions, and risk management. Red highlights mark incorrect or overconfident bullish claims; green highlights mark correct bearish signals aligned with the outcome.

Qwen3-VL-235B (zero-shot) — Outlook: BULLISH

> The 20-day candlestick chart displays a clear uptrend interrupted by two key periods of consolidation and minor reversal. [...] A decisive breakout occurs on Day 11, with a large green candle and strong volume, followed by a healthy pullback on Day 14 — signaling potential support. The subsequent rebound on Days 15–20 shows resilience: higher highs, higher lows, and steady volume. [...] The bullish interpretation is stronger: price consistently made higher highs after the Day 11 breakout, volume supported rallies, and corrections were shallow and short-lived. The market structure is not showing signs of

topping — no major distribution or bearish divergence. [...] STEADY_UP (55%), SELLOFF_THEN_STABILIZE (5%).

Structured advisory: Scenarios: STEADY_UP (55%), RALLY_THEN_FADE (25%), SIDEWAYS (15%), SELLOFF_THEN_STABILIZE (5%). Confidence: 0.70. Volatility: MODERATE. Key trigger: Price holds above $170 on Day 21. Risk factor: Breakdown below $170 with increasing volume. Max drawdown: -2.5%.

Claude Sonnet 3.7 (zero-shot) — Outlook: BULLISH

The stock has demonstrated a strong overall uptrend, moving from approximately $157 to $176, representing about a 12% gain over the 20-day period. [...] Recent bullish days (17–20) show moderately strong but decreasing volume, suggesting potential momentum fatigue — yet the conclusion remains: "a buy-the-dip mentality among market participants." [...] The pullback on Day 14 and Day 16 came with elevated volume, but bulls quickly reasserted control. [...] STEADY_UP (30%), SELLOFF_THEN_STABILIZE (10%).

Structured advisory: Scenarios: RALLY_THEN_FADE (40%), STEADY_UP (30%), SIDEWAYS (20%), SELLOFF_THEN_STABILIZE (10%). Confidence: 0.75. Volatility: MODERATE. Key trigger: Price action around $178–180 resistance. Risk factor: Exhaustion of buying momentum after extended rally. Max drawdown: -3.5%.

Hindsight DPO (4B) — Outlook: BEARISH ✓

The 20-day candlestick chart reveals a stock that has experienced a strong uptrend from approximately $155 to $175. [...] However, the most recent price action (Days 17–20) is concerning: despite closing higher on Day 19 and 20, the candles are becoming increasingly smaller, and the volume is not expanding on up days — a potential sign of weakening conviction. Day 16 is a large bearish candle that pulled back sharply after a rally, followed by a series of small-bodied candles with diminishing volume on Days 18–20. This suggests exhaustion or distribution at current levels. [...] The failure to break above $175 with conviction (no follow-through volume) may signal a RALLY_THEN_FADE. [...] RALLY_THEN_FADE (45%), SELLOFF_THEN_STABILIZE (35%).

Structured advisory: Scenarios: RALLY_THEN_FADE (45%), SELLOFF_THEN_STABILIZE (35%), SIDEWAYS (20%). Confidence: 0.65. Volatility: MODERATE. Key trigger: Stock opens above $175 on Day 1 with increasing volume. Risk factor: Positive catalyst reignites buying interest. Max drawdown: -2.5%.