```
            mentioned in ‘Rationale‘.

      - ICL Examples:
        - "‘SPX_Close_Slope‘␣>␣0␣&&␣‘VIX_Close_Slope
            ‘␣<␣0␣We␣have␣Market␣Confidence"
        - "‘GDP_QoQ‘␣Falling␣&&␣‘PMI‘␣<␣50␣We␣have␣
            an␣Economic␣Slowdown."

    5. Options Analysis:
      - Compare ‘OTM_Skew‘, ‘ATM_Skew‘, and ‘
          ITM_Skew‘ IV Skews: Assess differences to
          gauge market sentiment and directional
          bias using their ‘20Day_Moving_Averages‘.
      - Leverage IV spikes to capitalize on
          speculative directional trades.
      - Example: "Rising␣‘ATM_Skew_MA‘␣>␣0,␣market␣
          pricing␣up␣move,␣with␣stable␣HV␣supports␣a
          ␣LONG␣position,␣as␣it␣indicates␣growing␣
          upside␣expectations␣without␣excessive␣fear
          ."

    6. News Analysis:
      - Use ‘News_Sentiment‘ and ‘News_Impact_Score‘
           (1-3).
      - Only strong directional news (score = 3)
          should override other signals.
      - Medium news (score = 2) supports but does
          not lead.
      - Always check if news contradicts macro or
          technical trend.

    7. Performance Reflection and Strategic
        Adaptation:
      - If ‘Last_Strategy_Used_Data‘ is available:
            - Assess the outcome of the previous
                strategy by examining ‘last_returns‘
                 and the chosen ‘last_action‘.
            - Determine if the result aligns with
                the expectations outlined in the
                previous ‘Rationale‘.
            - Identify if the direction (LONG or
                SHORT) led to desirable or
                undesirable outcomes.
            - You must NOT reuse or copy the
                previous ‘Rationale‘. It is only
                context for reflection.
            - Summarize in 1-2 sentences whether the
                 previous strategy performed as
                expected.
            - Example: "The␣previous␣LONG␣strategy␣
                yielded␣positive␣returns,␣confirming
                ␣the␣bullish␣setup␣based␣on␣RSI␣and␣
                moving␣averages."
            - Do NOT include language or phrasing
                from the previous rationale.
      - Confidence assignment:
          - Assign a Likert score (1 to 3) to your
               ‘action_confidence‘:
              - 1: Low confidence; contradictory or
                  weak alignment across features.
              - 2: Moderate confidence; partial
                  alignment with moderate evidence.
              - 3: High confidence; strong
                  convergence across key features.
      - Feature Attribution:
          - Rank the importance of each major
               feature used in your current rationale
                using a Likert scale (1 to 3):
              - 1: Minimal contribution; not
                  required for the decision.
              - 2: Moderate contribution; relevant
                  but not critical.
              - 3: High contribution; pivotal to the
                   trading decision.

Output:
  action: Str. LONG or SHORT.
  action_confidence: int. Likert scale (1-3)
       confidence in the proposed ‘action‘, adjusted
       based on prior strategy outcome if ‘
       Last_Strategy_Used_Data‘ is available.
  explanation: >
    A concise rationale (max 350 words) justifying
        the proposed ‘action‘.
    Include:
      - The top 5 weighted features used in the
          decision, each labeled with its Likert
          importance (1-3).
        (e.g., "Stock_Data.Price.Close,␣Weight␣3,␣
            Technical_Analysis.RSI.Value,␣Weight␣1,␣
            Options_Data.ATM_Skew,␣Weight␣2")
      - A reflective assessment of ‘
          Last_Strategy_Used_Data‘, including
          whether the past ‘action‘ was successful
          and was it maintained given prior ‘
          Rationale‘.
  features_used:
    - feature: the features used from the prompt’s␣
        context.
      direction: LONG, SHORT, or NEUTRAL
      weight: A Likert score (1 to 3) described in
          Feature Attribution.
```

Appendix B
Analyst Prompt

The Analyst prompt used in Experiment 1 is presented in Listing 2, adapted from [10]. News corpora were anonymized prior to prompting.

Listing 2. Analyst Prompt

```
User_Context:
  Monthly_News_Articles_List: |
    "{articles_list}"

System_Context:
  Persona: Financial Market Analyst
  Instructions: |
    Extract the ‘Top 3‘ news factors influencing
         stock price movements from the ‘
         Monthly_News_Articles_List‘. Follow these
         steps:

    1. Rank the news by relevance to stock price
         movements:
       - Prioritize news related to significant
           financial or market impacts (e.g.,
           acquisitions, partnerships, guidance
           revisions).
       - Weigh industry trends, macroeconomic
           influences, and analyst ratings based on
           their expected effect on the company
           valuation.
       - News with broad or long-term implications
           ranks higher.

    2. Summarize content into key factors and
         corporate events affecting stock prices,
         using concise language and causal
         relationships.

    3. For each factor, assign:
       - ‘Sentiment‘: +1 for positive, -1 for
           negative, 0 for neutral or mixed
       - ‘Market_Impact_Score‘: Likert scale from 1
           to 3, where:
         - 1 = minimal relevance
         - 2 = moderate influence
         - 3 = high impact driver

    Examples of factors influencing stock prices
         include:
      - Strategic partnerships or competitor
          activity.
      - Industry trends or macroeconomic influences.
      - Product launches or market expansions.
      - Analyst ratings, significant stock price
          moves, or expectations.
      - Corporate events: guidance revisions,
          acquisitions, contracts, splits,
          repurchases, dividends.

    Example:
```

Algorithm 1: Expert Trade Heuristic

```
Data: Time-indexed price series
Result: Trade action: LONG (1) or SHORT (0)
foreach date $t$ in dataset do
    $P_t \leftarrow \text{Close}(t)$;
    $r^{(10)} \leftarrow \frac{P_{t+10}}{P_t} - 1$, $r^{(20)} \leftarrow \frac{P_{t+20}}{P_t} - 1$;
    $r^{\text{weighted}} \leftarrow 0.4 \cdot r^{(10)} + 0.6 \cdot r^{(20)}$;
    if $r^{weighted} >= 0$ then
        Action ← LONG    (Trade_Action = 1);
    else
        Action ← SHORT    (Trade_Action = 0);
```

```
        'A␣major␣tech␣company␣partners␣with␣a␣leading␣
            automotive␣firm␣for␣EV␣battery␣innovation.
            ␣Analysts␣predict␣this␣could␣boost␣
            revenues␣significantly.'

        Ranked Factors:
          1. factor: Strategic partnership in EV
               battery technology expected to increase
               revenue.
             sentiment: +1
             market_impact: 3
          2. factor: Positive sentiment driven by
               projected long-term gains.
             sentiment: +1
             market_impact: 2
          3. factor: Growing demand for EV technology
               anticipated to support future earnings.
             sentiment: +1
             market_impact: 2

  Output:
    factors:
      - factor: str. Summary of the news item. Max 70
          words.
      - sentiment: int. One of Positive +1, Negative
          -1, or Neutral 0
      - market_impact: int. Likert scale 1 to 3
```

Appendix C
Algorithms

The labeling algorithm emulates expert trading behavior by deliberately leveraging future return information to assign proxy trade actions in hindsight. This approach offers a cost-effective and scalable addition to manual annotation, capturing the general direction an informed trader might take. These synthetic labels are then provided to the LLM, along with a smaller set of HITL annotated examples.

Appendix D
Dataset

Market Data

This market data ($\mathcal{S}_{\text{mk}}$) included OHLCV price series as well as macro-level indicators and forward-looking sentiment signals. Specifically, it comprised:

- Daily returns of the S&P 500 Index (SPX) and NASDAQ-100 Index (NDX). These are market and sector indices,
- Implied Volatility (IV) and Historical Volatility (HV) metrics, derived from the stock's derivatives,
- The CBOE Volatility Index (VIX) as a proxy for market fear and option market expectations,
- *Weekly Past Returns*, which record the percentage change over the past four weekly intervals. The four-week span was selected empirically to align with the model's monthly strategy generation frequency.

These features help in modeling short-term market dynamics.

Fundamental Data

Fundamental data ($\mathcal{S}_{\text{fund}}$) has firm-level fundamentals and macroeconomic indicators. Macroeconomic variables provided contextual narrative for interpreting observed signals, and supporting regime identification [8], [9]. This set covered:

- **Liquidity ratios:** Current Ratio, Quick Ratio;
- **Leverage and coverage:** Debt-to-Equity, Interest Coverage;
- **Profitability metrics:** Gross Margin, Operating Margin, Return on Equity (ROE), Return on Assets (ROA);
- **Valuation:** Price-to-Earnings (P/E), Price-to-Book (P/B), Enterprise Value (EV), and Earnings Before Interest, Taxes, Depreciation, and Amortization (EBITDA).
- **Growth:** Revenue and Earnings Growth;
- **Macroeconomic indicators:** Gross Domestic Product (GDP), Purchasing Managers' Index (PMI), Producer Price Index (PPI), Consumer Confidence Index (CCI), U.S. 10-Year Treasury Yield, and the 10Y–2Y yield curve slope.

To enhance temporal abstraction, all variables were computed as quarter-over-quarter (QoQ) or year-over-year (YoY) percentage changes. It is critical to take first-order dynamics as LLMs can recall absolute numbers for economic details, allowing look-ahead bias in the backtests [20].

Analytics

Technical indicators ($\mathcal{S}_{\text{an}}$) were computed over rolling 20-day windows using the open-source TA-Lib[7] library. These features include:

- Simple Moving Averages (SMA) over 20, 50, 100, 200 trading-day horizons,
- Relative Strength Index (RSI),
- Average True Range (ATR) for volatility,
- Moving Average Convergence Divergence (MACD) with its signal line and derived strength,
- Volume-Weighted Average Price (VWAP) as a reference anchor for intraday valuations.

Each indicator was extended with slope and z-score to assist the LLM in capturing directional shifts and the statistical significance of deviations. These technical indicators are widely used in trading practice and academic research [18].

Alternative Data

Structured representations of financial news headlines ($\mathcal{S}_{\text{alt}}$) were extracted using a large language model (LLM), which anonymized and synthesized the content into latent factors. Following the LLMFactor methodology [10], each news item was distilled into 2–5 interpretable factors, capturing macroeconomic and firm-specific signals.

[7] https://ta-lib.org/

Instrument	**Paper SR**	**SR (±σ) [p-value]**	**MDD (±σ)**
AB InBev	0.187	**1.21 (0.30) [0.00]**	0.18 (0.08)
Alibaba	0.021	**0.06 (0.02) [0.00]**	0.09 (0.01)
Amazon	0.419	0.39 (0.45) [0.85]	0.30 (0.09)
Apple	1.424	1.19 (0.55) [0.22]	0.29 (0.09)
Baidu	0.080	**0.20 (0.17) [0.00]**	0.36 (0.09)
CCB	0.202	**0.33 (0.25) [0.04]**	0.24 (0.14)
Coca Cola	1.068	1.07 (0.53) [0.50]	0.25 (0.04)
Dow Jones	0.684	0.70 (0.30) [0.91]	0.25 (0.05)
ExxonMobil	0.098	0.10 (0.35) [0.91]	0.34 (0.08)
FTSE 100	0.103	**0.50 (0.23) [0.00]**	0.31 (0.08)
Google	0.227	**-0.54 (0.59) [0.00]**	0.43 (0.13)
HSBC	0.011	**0.38 (0.17) [0.00]**	0.29 (0.05)
JPMorgan Chase	0.722	0.72 (0.31) [0.98]	0.26 (0.06)
Kirin	0.852	0.85 (0.42) [0.99]	0.39 (0.07)
Meta	0.151	**0.63 (0.61) [0.01]**	0.45 (0.27)
Microsoft	0.987	0.70 (1.00) [0.38]	0.28 (0.16)
NASDAQ 100	0.845	0.85 (0.35) [1.00]	0.16 (0.05)
Nikkei 225	0.019	**0.26 (0.29) [0.02]**	0.29 (0.07)
Nokia	-0.094	**0.07 (0.24) [0.00]**	0.57 (0.15)
PetroChina	0.156	0.22 (0.29) [0.29]	0.67 (0.00)
Philips	0.675	**1.40 (0.50) [0.00]**	0.25 (0.03)
S&P 500	0.834	0.83 (0.25) [1.00]	0.14 (0.04)
Shell	0.425	0.42 (0.37) [0.95]	0.51 (0.05)
Siemens	0.426	0.39 (0.23) [0.43]	0.26 (0.12)
Sony	0.424	0.42 (0.36) [0.97]	0.16 (0.04)
Tesla	0.621	0.48 (0.41) [0.29]	0.52 (0.09)
Tencent	-0.198	-0.19 (0.33) [0.98]	0.10 (0.09)
Toyota	0.304	0.36 (0.27) [0.37]	0.45 (0.10)
Volkswagen	0.216	**0.45 (0.18) [0.00]**	0.48 (0.09)

TABLE XI
REPLICATION METRICS FOR [3]

To mitigate memorization and data leakage risks, named entities and dates were anonymized (e.g., “Tesla” becomes “the Company”).

APPENDIX E
REPLICATED BENCHMARK METRICS

We report the replicated benchmark metrics in Appendix E for the assets used in [3]. We include the mean SR and MDD, each averaged across 25 runs with standard deviation σ.

For the SR, we conduct a two-sided one-sample t-test to assess whether the metric is significantly different from the published value. The null hypothesis H_0 assumes equivalence, i.e. $H_0 : \mu_{\text{SR}} = \text{SR}_{\text{paper}}$.

Since this is a replication test, failing to reject H_0 indicates successful replication. p-values are computed only for SR; other metrics are reported without significance testing.

All assets have been successfully replicated within acceptable bounds, with exceptions highlighted in bold. Notably, GOOGL, one of the stocks included in our test environment, exhibited a statistically significant deviation from the original benchmark, with a p-value below 0.05.