TABLE I
PROMPT VERSIONS USED IN EXPERIMENT 1

Prompt	Description
P0	Baseline prompt containing only static technical indicators and price features.
P1	Augmented P0 with selected features and instructions.
P2	P1 extended with ICM, incorporating prior strategy.
P3	P2 extended with instruction decomposition and CoT reasoning across six structured signal groups.
P4	P3 enriched with macroeconomic and firm-specific news-derived directional signal.

aged verbatim reuse from the KB or ICM while preserving exploratory diversity.

For strategy generation, temperature was set to 0 with fixed seed 49 for reproducibility. Strategies were produced on a monthly cadence (20 trading days), aligning with common guidance/rebalancing cycles and remaining tractable given LLM inference cost.

At most three refinement iterations were permitted ($T \leq 3$). Convergence was declared when the regret $\mathcal{R}(T)$ approached zero or when the SR exceeded the initial threshold $\max\{V_{\text{baseline}}, 0.8\}$. The procedure was repeated five times with discretionary HITL adjustments between runs. The iteration count balanced methodological tractability against computational cost.

All technical indicators used a 20-trading-day rolling window with standard *TA-Lib* defaults (e.g., 14-day RSI).

a) Quantitative Metrics: We evaluate LLM-generated strategies using three complementary metrics: risk-adjusted returns, model confidence, and model uncertainty.

The SR serves as the core risk-adjusted returns metric:

$$\text{SR} = \frac{\mathbb{E}[R_t - R_f]}{\sigma_R}. \tag{2}$$

where R_t is the portfolio return, R_f is the risk-free rate, and σ_R is the return volatility. SR also serves as a proxy for the LLM's financial reasoning [12], [14]. To ensure comparability across different periods for the daily returns, we annualize the SR to 252 trading days per year: Annualized SR $= \text{SR} \cdot \sqrt{252}$.

As a proxy for prompt quality we compute the Perplexity (PPL) [21] over the LLM-generated strategies:

$$\text{PPL} = \exp\left(-\frac{1}{N}\sum_{t=1}^{N}\log p(w_t \mid w_{<t})\right). \tag{3}$$

where $p(w_t \mid w_{<t})$ denotes the conditional token probability. Lower values indicate higher quality.

To complement this, we report token-level entropy H_{LLM}, approximated using top-k distributions:

$$H_{\text{LLM}} = \frac{1}{N}\sum_{t=1}^{N}\left(\sum_{v\in V_k} -p_t(v)\log p_t(v) - p_{\text{tail},t}\log p_{\text{tail},t}\right). \tag{4}$$

where V_k denotes the top-k token set and $p_{\text{tail},t}$ represents the unobserved probability mass [22]. In our experiments, $k = 5$.

TABLE II
EXPERT RUBRIC FOR SCORING LLM RATIONALES

Criterion	1	2	3
Rationale	Flawed	Partial	Sound
Fidelity	Unrealistic	Plausible	Professional
Safety	Ignored	Mentioned	Addressed

Lower entropy indicates greater decisiveness, whereas higher values suggest uncertainty.

Together, PPL and H_{LLM} enable a measurement of prompt quality and strategy confidence.

b) Qualitative Evaluation: Qualitative assessment was conducted via an Expert Review Score (ERS), a human-grounded rubric evaluating LLM-generated trading rationales along three dimensions: economic rationale, domain fidelity, and trade safety (risk awareness). Each dimension was scored on a 3-point ordinal scale {1 = poor, 2 = average, 3 = good}, based on the rubric shown in Table II.

The review process followed a similar setup to that of [23], involving ten participants: five senior finance professionals and five retail traders (or professionals in the industry who do not actively trade). Each reviewer evaluated anonymized data for three instruments over one year, including price data, fundamental and macroeconomic metrics, and firm-level news headlines.

Before reviewing the LLM rationale, expert participants made their own directional prediction (`LONG`/`SHORT`) to activate their internal domain models. They then reviewed the LLM's reasoning and scored it using the rubric. Each session concluded with a 60-minute structured discussion to elicit LLM critiques and identify exemplars to use. Surveys took approximately 15 minutes to complete. All scores were normalized to a 1–3 range.

C. Experiment 2: LLM-Guided RL

This experiment addressed Objective 2 by incorporating the LLM guidance within an RL framework.

1) Data and Feature Engineering: The LLM outputs from Experiment 1 were reused. The RL agent adopted the DDQN configuration of [3], with a single LLM-derived interaction term τ as innovation to the observation space. This feature consisted of:

- **Signal Direction** ($\text{dir}(\pi^g)$): The discrete directional recommendation from the LLM. Zero represents `SHORT` and one `LONG`.
- **Signal Strength** ($\text{str}(\pi^g)$): The LLM's entropy-adjusted confidence score as a Likert-3 score.

The interaction term was defined as

$$\tau = \text{dir}(\pi^g) \cdot \text{str}(\pi^g), \tag{5}$$

where $\text{dir}(\pi^g)$ was remapped from $\{0, 1\}$ to $\{-1, 1\}$ to enable the interaction.

The LLM's signal strength was derived from the normalized LLM's confidence score:

$$\mu_{\text{conf}} = \frac{\text{Likert}}{3}, \tag{6}$$

and adjusted using entropy-based certainty:

$$C = \varepsilon + (1 - \varepsilon)(1 - H), \tag{7}$$

where $H \in [0, 1]$ is the normalized entropy of the LLM output, and $\varepsilon = 0.01$ ensures numerical stability. The final strength term is:

$$\mathrm{str}(\pi^g) = \mu_{\text{conf}} \cdot C. \tag{8}$$

This entropy-adjusted confidence follows the approach of [24], providing a soft weighting of the LLM's signal by its certainty.

The interaction term τ was selected empirically. Initial variants used direction only ($\mathrm{str}(\textit{dir})$), followed by LLM's confidence ($\mathrm{str}(\pi^g)$) and direction. The final form was chosen based on empirical performance and compatibility with DDQN's continuous normalized input space [3].

2) LLM+RL Hybrid Architecture: The baseline DDQN agent is augmented by the Strategist Agent and Analyst Agent, which produce monthly strategies for the stock's behavior. For practical reasons, outputs from the LLM were precomputed per instrument and fixed throughout training.

3) Training and Parameters: Hyperparameters mirror [3] and the LLM settings follow those in Experiment 1. Training was conducted over 25 runs × 50 episodes per instrument using an NVIDIA RTX 3050, with each equity trained for 3 hours.

To ensure comparability with the benchmark [3], we replicated all baseline metrics within acceptable statistical bounds.

4) Evaluation Metrics: Two measures were considered:

- **SR:** Same as Experiment 1 see Eq. (2).
- **MDD:** Captures the largest observed loss from a historical peak to a subsequent trough:

$$\mathrm{MDD} = \frac{P_{\text{peak}} - P_{\text{low}}}{P_{\text{peak}}}. \tag{9}$$

  where P_{peak} is the highest portfolio value observed before the largest drop, and P_{low} is the lowest value reached before a new peak is established. Lower values indicate stronger downside protection.

These metrics together assess whether LLM-guided RL agents can adapt to different equities without changing the core architecture.

III. Results and Discussion

A. Experiment 1 Results

This section presents empirical results across the baseline (P0) and four prompt versions (P1–P4) from Table I, addressing Objective 1. We evaluated their impact on SR, PPL, and H_{LLM}. From P4 onward, qualitative evaluation was incorporated through ERS, introduced once the prompt design had stabilized. All backtests were conducted over 2018–2020 to ensure comparability with the RL's OOS results. Statistical significance was assessed using two-tailed *t*-tests across 25 runs per ticker, with hypotheses $H_0 : \mu_{\text{P4}} = \mu_{\text{P1}}$. All runs were executed at a sampling temperature of 0.7 to capture variance, while the reported metrics correspond to the deterministic setting (temperature 0) with fixed random seeds for reproducibility.

TABLE III
Sharpe Ratio Across Prompts and Benchmark

Ticker	P0	P1	P2	P3	P4	BM
AAPL	1.13	1.09	1.07	1.07	**2.09**	1.27
AMZN	0.51	0.35	0.38	0.63	**0.84**	0.21
GOOGL	0.34	0.26	0.52	0.52	**1.12**	0.19
META	0.60	-0.06	-0.28	0.30	**0.77**	0.63
MSFT	0.36	1.07	1.11	**1.31**	0.50	1.17
TSLA	0.34	0.71	0.75	0.43	**0.79**	0.67
Mean	0.55	0.57	0.59	0.71	**1.02**	0.69

TABLE IV
Perplexity Across Prompts

Ticker	P0	P1	P2	P3	P4
AAPL	1.44	1.85	**1.31**	1.55	1.44
AMZN	1.51	1.74	1.35	1.68	**1.31**
GOOGL	1.56	1.77	1.49	1.78	**1.33**
META	1.47	1.73	**1.31**	1.39	1.38
MSFT	1.43	1.83	1.44	1.49	**1.24**
TSLA	1.46	1.77	1.50	1.63	**1.39**
Mean	1.48	1.78	1.40	1.59	**1.35**

Tables III–V summarize the results across prompt versions relative to the benchmark (BM). Prompt 0, which relied solely on static technical features, outperformed Prompt 1 primarily because all equities exhibited upward trends during the OOS period. Prompt 1 yielded the weakest performance, with the lowest SR across most equities and the highest PPL and entropy, indicating that the LLM was unable to exploit the additional information when presented in an isolated context. Prompt 2 incorporated ICM, producing moderate gains in SR (mean 0.59) and suggesting improved confidence through reflection. Prompt 3 introduced decomposed instructions, eliciting CoT, and outperformed the benchmark with a mean SR of 0.71. Prompt 4 further included unstructured news signals and achieved the highest mean SR (1.02), lowest PPL and entropy, and showed higher confidence particularly on sentiment-sensitive tickers such as TSLA. Based on the p-values in Table VII, the improvements were statistically significant for SR and entropy, while the changes in PPL were comparatively weaker.

Expert evaluation of Prompt 4 confirmed its effectiveness. Reviewers rated the LLM's rationale highly (mean 2.7 out of 3), highlighting its ability to integrate valuation, sentiment, and analytics.

Fidelity received a slightly lower score (mean 2.65), with critiques focused on inconsistent thresholding. For instance, one reviewer noted, *"Calling RSI near* 40 *'oversold' is debatable,"* requiring refinements in numerical phrasing.

Feedback varied by background: buy-side professionals emphasized transparency in feature weighting, whereas retail reviewers focused on technical and macro signals. All com-

TABLE V
ENTROPY ACROSS PROMPTS

Ticker	P0	P1	P2	P3	P4
AAPL	0.66	0.70	0.67	**0.66**	0.69
AMZN	0.69	0.69	0.69	0.69	**0.67**
GOOGL	0.67	0.67	0.67	0.70	**0.66**
META	0.68	**0.66**	0.70	0.73	0.67
MSFT	0.65	0.66	0.68	0.72	**0.65**
TSLA	0.67	0.68	0.70	0.74	**0.65**
Mean	0.67	0.68	0.69	0.71	**0.67**

TABLE VI
EXPERT REVIEWER SCORES FOR PROMPT 4

Dimension	ERS (1–3)
Rationale	2.70
Fidelity	2.65
Safety	2.80

mented on the lack of a neutral or hold signal, which was done to align with [3].

Overall, results validated Prompt 4's modular design and market narrative awareness. It outperformed earlier prompts and was selected as the global policy generation prompt for the LLM–RL hybrid in Experiment 2.

The computational costs of Experiment 1 are summarized in Table VIII. The overall cost for a single run with each prompt was approximately $36, increasing to about $150 when the writer–judge loop was included. When accounting for additional trials and development, the cumulative cost amounted to $345. Inference time ranged from approximately 1.5 to 2 hours per asset and prompt version.

B. Experiment 2 Results

This experiment addressed Objective 2 by comparing three agent architectures: (i) the benchmark RL-only [3], (ii) the best-performing LLM prompt from Experiment 1, and (iii) a hybrid LLM+RL agent. All agents were trained in identical environments.

To determine whether the hybrid agent outperformed the benchmark, we conducted two-sided paired t-tests on the SR across 25 runs for each stock. The null hypothesis H_0 assumed no difference in mean performance: $H_0 : \mu_{\text{LLM+RL}} = \mu_{\text{RL-only}}$. All resulting p-values were below 0.05, indicating statistically significant improvements.

Results in Table IX confirm that the LLM+RL agent outperformed the RL-only baseline in four out of six assets.

AAPL and META did not show consistent individual outperformance. Fig. 2 illustrates AAPL's trading behavior during one episode. The top panel plots price, technical indicators, and trades: hollow triangles mark RL trades; filled arrows show LLM monthly guidance. The LLM issued sparse but confident signals (strength > 0.6), often aligned with technical points of interest (e.g., MA interactions). In contrast, the RL agent frequently mistimed entries and exits.

TABLE VII
P4 VS. P1 SIGNIFICANCE OF METRIC CHANGES

Metric	*t*-test p-value
Entropy	2.29×10^{-4}
Perplexity	7.25×10^{-2}
Sharpe Ratio	2.3×10^{-5}

TABLE VIII
TOKEN USAGE AND COSTS

Prompt	Mean Tokens	Mean Cost($)	Total Tokens	Total Cost($)
v0	663	$0.00020	2.0×10^6	$1.19
v1	1,760	$0.00043	3.5×10^6	$5.62
v2	2,240	$0.00051	4.5×10^6	$6.48
v3	3,300	$0.00067	6.6×10^6	$8.75
v4	8,300	$0.00150	1.6×10^7	$21.60

From December 2018 to January 2019, the RL agent oscillated between LONG and SHORT positions with punishing results and despite receiving strong signals from the LLM. The LLM issued high-confidence guidance for a SHORT in December followed by a LONG in January, both with signal strengths exceeding 0.8. Regardless, the RL agent held a LONG position throughout the decline.

As shown in Fig. 5, the DDQN assigns lower Q-values to SHORT actions, indicating limited confidence. This follows from lower-bound constraints (used to cap leverage) that created an asymmetric return function by triggering buy-to-cover after price increases, reducing portfolio value and subsequent SHORT exposure. Also, the selected equity universe has positive historical drift, which raises average prices, with limited opportunity to capture SHORT returns. Together these features lower the expected return of a SHORT and discourage sustained SHORT positions [3].

The bottom panel confirms that the LLM maintained high confidence near key inflection points, and reduced conviction when trends have persisted (possibly awaiting a reversal from its training corpus). However, the RL agent didn't fully exploit these signals due to the underlying RL architecture, which remained fixed for the purposes of this experiment.

Fig. 3 illustrates the evolution of the SR for AAPL throughout the training episodes. The hybrid LLM+RL agent (orange line) outperformed the baseline RL agent (blue line) in both mean Sharpe and stability, as reflected in the narrower shaded confidence intervals. The LLM's SR is shown for reference (black dashed line).

Figs. 4 and 5 show Q-values for `LONG` and `SHORT` actions respectively, with y-axis clipped to $[-0.03, 0.03]$ to highlight late-episode convergence. Early training was noisy for both agents. The LLM+RL agent converged faster with lower variance. Although Q-value separation rarely exceeded 0.01, the hybrid showed slightly stronger directional signals. These gains emerged without modifying the DDQN or imposing reward shaping, thus isolating the effect of the LLM's guidance. The narrow Q-range stems from the RL baseline design.