Alpha-R1: Alpha Screening with LLM Reasoning via Reinforcement Learning

Anonymous ACL submission

Abstract

Signal decay and regime shifts pose recurring challenges for data-driven investment strategies in non-stationary markets. Conventional time-series and machine learning approaches, which rely primarily on historical correlations, often struggle to generalize when the economic environment changes. While large language models (LLMs) offer strong capabilities for processing unstructured information, their potential to support quantitative factor screening through explicit economic reasoning remains underexplored. Existing factor-based methods typically reduce alphas to numerical time series, overlooking the semantic rationale that determines when a factor is economically relevant.

We propose Alpha-R1, an 8B-parameter reasoning model trained via reinforcement learning for context-aware alpha screening. Alpha-R1 reasons over factor logic and real-time news to evaluate alpha relevance under changing market conditions, selectively activating or deactivating factors based on contextual consistency. Empirical results across multiple asset pools from markets across different countries show that Alpha-R1 consistently outperforms benchmark strategies and exhibits improved robustness to alpha decay.

1 Introduction

Factor investing has remained a cornerstone of asset management since the seminal work of Fama and French (Fama and French, 1993), evolving from foundational CAPM (Sharpe, 1964) and multi-factor models (Carhart, 1997) to the modern high-dimensional "Factor Zoo" (Feng et al., 2020). Simultaneously, natural language processing (NLP) has enabled the extraction of sentiment signals from unstructured news and financial reports (Lopez-Lira and Tang, 2023). This field has shifted toward domain-specific foundation models, with examples like BloombergGPT (Wu et al., 2023) and FinGPT (Liu et al., 2023) validating the efficacy of pre-training and fine-tuning (e.g., LoRA (Hu et al., 2022)) for specialized tasks like sentiment analysis and entity recognition (Xie et al., 2023; Zhang et al., 2023a; Lu et al., 2023; Zhang et al., 2023b). However, a key gap remains that traditional numerical and textual signals are often treated as separate modalities via static rules rather than within a unified framework that captures their semantic interactions in dynamic markets (Wu et al., 2023; Liu et al., 2023).

The financial domain's inherent non-stationarity and noise (Feng et al., 2020; He et al., 2025) necessitate adaptive reasoning to navigate uncertainty, such as assessing factors under shifting macroeconomic conditions. These demands contrast with the deterministic nature of typical LLM benchmarks (Hendrycks et al., 2021). While sparsity-based machine learning (e.g., Lasso) exists (Freyberger et al., 2020), it often suffers from poor interpretability and instability during regime changes (Freyberger et al., 2020), whereas general-purpose LLMs often lack alignment with financial principles for transparent decision traces (Tatsat and Shater, 2025).

Current LLM research in finance has largely focused on factor mining and information extraction (Cheng et al., 2024; Wang et al., 2024). Generative frameworks like Chain-of-Alpha (Cao, 2025), Alpha-GPT (Wang et al., 2025; Yuan et al., 2024), R&D-Agent-Quant (Li et al., 2025), and AlphaAgent (Tang et al., 2025) explore iterative refinement and multi-agent systems to bridge hypothesis generation and code implementation, while others utilize multimodal data for trading signals (Kou et al., 2025) and efficient evaluation (Ding et al., 2025). Nevertheless, a gap remains between broad mining and the path-dependent reasoning required for systematic factor screening. While efforts like Trading-R1 (Xiao et al., 2025a), Fin-R1 (Liu et al., 2025), and FinO1 (Qian et al., 2025) enhance financial reasoning, screening decisions remain highly

context-dependent.

To address these structural limitations, we propose Alpha-R1, a dynamic investment framework centered on a specialized reasoning model trained via reinforcement learning (Shao et al., 2024; Guo et al., 2025). Distinct from generic LLM applications or purely agentic frameworks (Hong et al., 2024; Xiao et al., 2025b), Alpha-R1 serves as the system's cognitive core, designed to support the sequential reasoning required for dynamic factor screening. It inductively reasons over heterogeneous market information and real-time news to assess the economic relevance of candidate factors, constructing portfolios that align with prevailing market conditions. Furthermore, Alpha-R1 attributes return sources in a structured manner, providing explanations for selection decisions and addressing the inherent opacity of traditional quantitative models.

Our primary contributions are as follows. First, we bridge static quantitative models and dynamic markets with a practical framework. By synthesizing heterogeneous information (e.g., macro indicators, news), it enables regime-aware factor screening and dynamically adjusts portfolio exposure based on semantic alignment between factor rationales and market conditions. Second, we design a specialized reasoning core by adapting reinforcement learning from human feedback (RLHF) to finance, replacing subjective preferences with objective rewards based on realized market performance. This aligns the reasoning process with trading objectives and supports sequential decision-making under uncertainty. Finally, extensive backtesting across diverse global asset pools demonstrates that Alpha-R1 consistently outperforms state-of-the-art benchmarks and traditional factor strategies, exhibiting robustness to alpha decay and superior risk-adjusted performance.

2 Related Work

2.1 LLMs in Quantitative Trading

Foundation Models and Adaptation. Initial adaptation focused on domain-specific pre-training, established by BloombergGPT (Wu et al., 2023), and instruction tuning via efficient methods like FinGPT (Liu et al., 2023) and Instruct-FinGPT (Zhang et al., 2023a). Recent work has explored Chinese financial benchmarks such as BBT-Fin (Lu et al., 2023) and XuanYuan 2.0 (Zhang et al., 2023b), alongside comprehensive evaluation frameworks like PIXIU (Xie et al., 2023). These models provide the linguistic foundation for complex quantitative tasks.

Alpha Factor Generation. LLMs have catalyzed a shift in alpha factor mining, moving from signal generation to sophisticated agent-based frameworks (Tang et al., 2025; Cao, 2025). Notable trajectories include mitigating alpha decay through iterative refinement (AlphaAgent (Tang et al., 2025)), dual-chain architectures for backtesting feedback (Chain-of-Alpha (Cao, 2025)), and two-stage generative neural networks (AlphaForge (Shi et al., 2025)). Furthermore, systems explore human-AI collaboration to bridge signal extraction and code implementation (Wang et al., 2025; Yuan et al., 2024).

Quant Agents. Research has expanded toward holistic agents managing the investment lifecycle, from return forecasting with unstructured news flow (Guo and Hauptmann, 2024) to multi-agent orchestration for risk management (Kou et al., 2025; Cheng et al., 2024; Hong et al., 2024). Comprehensive ecosystems like TradingGPT (Li et al., 2023), FinMem (Li et al., 2024), and FinAgent (Cao et al., 2025; Zhang et al., 2024) employ multimodal perception (Kou et al., 2025) and collaborative workflows (e.g., TradingAgents (Xiao et al., 2025b)) to autonomously navigate trading activities. These agents often integrate self-reflection to mitigate common hallucination issues (Koa et al., 2024; Ji et al., 2023). Unlike these generalist agents, Alpha-R1 focuses on the specific challenge of factor selection in non-stationary markets, leveraging semantic reasoning to bridge the gap between static factor definitions and dynamic market regimes.

2.2 Methods for Screening Alpha Factors

Sparsity-Driven and Non-Linear Selection. Direct machine learning approaches to the Factor Zoo include sparsity-inducing regularization like Lasso (Tibshirani, 1996; Mai et al., 2024), which yields stability and accuracy compared to unregularized methods (Mai et al., 2024). Alternatively, tree-based ensembles (XGBoost, LightGBM) and regression trees demonstrate exceptional performance in capturing nonlinear interactions for return prediction (Gu et al., 2020; Mai et al., 2024). Deep learning techniques further offer dimension reduction through autoencoders (Mai et al., 2024) or hybrid frameworks like CPCA (Mai et al., 2024) to maintain interpretability. However, the black-box

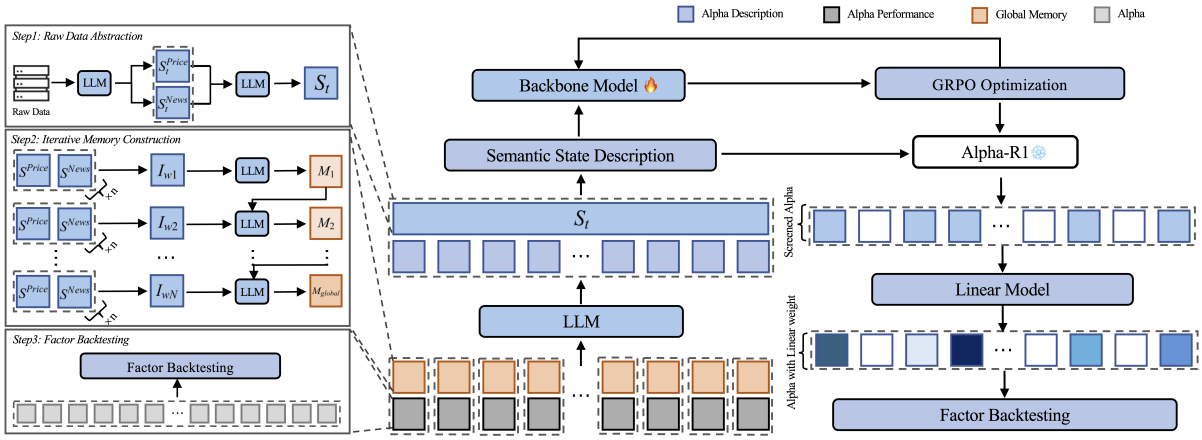


Figure 1: Alpha-R1 Framework Overview. The pipeline follows a sequential logic: (a) Preparation: Abstracting raw technical indicators and financial news into atomic textual units to construct a global historical memory (M_{global}), coupled with systematic historical factor backtesting; (b) Semantic Construction: Mapping quantitative performance metrics into structured semantic factor profiles (α_{des}) and synthesizing dynamic market states (S_t); (c) Reasoning & Optimization: Performing context-aware alpha screening via a reasoning core that evaluates α_{des} against S_t, with the policy iteratively refined through GRPO.

nature of these models often obscures economic rationale, a limitation Alpha-R1 addresses through explicit semantic reasoning.

2.3 Reinforcement Learning for LLMs

Alignment and Reasoning Optimization. Reinforcement learning is key to aligning LLMs with domain objectives (Kaufmann et al., 2025). While RLHF via PPO (Ouyang et al., 2022; Schulman et al., 2017) is standard, memory costs and reward hacking (Casper et al., 2023; Lambert et al., 2025) have spurred alternatives like DPO (Rafailov et al., 2023) and SimPO (Meng et al., 2024). For reasoning tasks, Group Relative Policy Optimization (GRPO) (Shao et al., 2024) offers a robust critic-free method to foster self-evolving reasoning chains (Guo et al., 2025). This paradigm is expanding in finance with Trading-R1 (Xiao et al., 2025a), Fin-R1 (Liu et al., 2025), and FinO1 (Qian et al., 2025). We leverage this architecture in Alpha-R1 to optimize factor screening against sparse market signals.

3 Methodology

We propose Alpha-R1, a novel framework that integrates semantic reasoning with reinforcement learning to dynamically select alpha factors in non-stationary markets. As illustrated in Figure 1, the pipeline consists of three sequential stages: (a) Preparation, (b) Semantic Construction, and (c) Reasoning & Optimization.

3.1 Data, Memory, and Factor Baselines

We establish the foundational data structures, historical context, and quantitative benchmarks through a systematic process.

3.1.1 Raw Data Abstraction

We first transform heterogeneous raw data into structured textual atomic units. At each time step t, we employ Large Language Models to summarize raw inputs into two complementary market descriptors. Figure 2 illustrates a generated Price Market Description sample. Additional examples, including News Market Summaries and US market data, are detailed in Appendix A.

- Price Market Description (S_t^{price}): Information from technical indicators, trading volume, and sector rotation patterns.
- News Market Description (S_t^{news}): Encodes information from financial news and macroeconomic announcements to capture prevailing market sentiment.

3.1.2 Iterative Memory Construction

We employ an iterative memory construction pipeline to capture long-term market context. This process aggregates the atomic textual units into a coherent historical narrative. Let $I_w = \{S_t^{\text{price}}, S_t^{\text{news}}\}_{t \in w}$ denote the set of market descriptions within week w. The weekly market summary

[Sample 1] Price Market Description ($S^{\text{price}}_{20240409}$)

90-Day Market Trend Review: Over the past 90 days, the major indices (Shanghai Composite, ChiNext, STAR 50) exhibited a synchronized cycle of "Rise (Late Nov–Mid Dec) → Adjustment → Rebound (Late Jan–Feb) → Decline (Mar–Apr)". Significant divergence in volatility was observed: the STAR 50 had the highest amplitude (∼21%), followed by ChiNext (∼17%) and Shanghai Composite (∼9%). Currently, all indices have closed below their initial levels from 90 days ago, indicating a downward shift in the market center of gravity. Volume analysis shows that growth indices (STAR 50, ChiNext) saw sustained volume amplification during the recent decline, indicating active capital turnover.
Intraday Characteristics (2024-04-08): The indices displayed a highly synchronized "Morning Spike → Mid-day Consolidation → Late-session Rally" pattern: (1) Morning (9:30–9:45): Rapid surge to daily highs followed by a swift retracement; volume peaked at the open. (2) Mid-day (9:45–14:30): Range-bound oscillation with shrinking volume; ChiNext adjusted the most. (3) Late-session (14:30–15:00): Collective rally. The Shanghai Composite and STAR 50 showed significant gains with expanding volume, closing near daily highs (+1.58% and +1.72% respectively).

Figure 2: Semantic description generated from price market data processed by a Claude 3.7 Sonnet.

M_w is updated recursively using a large language model:

$$M_w = F^{\text{mem}}_{\text{LLM}}(I_w \oplus M_{w-1}), \tag{1}$$

where $\oplus$ denotes textual concatenation and $w \in \{1, \ldots, W\}$ represents the week index. After iterating through the historical observation period of W weeks, we obtain the comprehensive historical market description as $M_{\text{global}} = M_W$. This encapsulates the structural evolution and regime shifts of the market, forming the long-term historical memory required for the subsequent semantic profiling.

3.1.3 Factor Backtesting

To establish the ground truth for factor behavior, we perform a backtest on the entire factor pool $\mathcal{U}$ over the historical window. For each factor i, we obtain a quantitative performance vector P_i, which includes key metrics such as returns, volatility, and decay characteristics. This dataset serves as the objective basis for linking market memory with factor effectiveness.

3.2 Profiling and State Description

Building on the prepared data foundation and long-term market memory, we construct the semantic state representations required for decision-making.

3.2.1 Factor Semantic Descriptions

This stage maps quantitative signals into structured semantic representations. We combine the global market description from the historical observation period, M_{global}, with the factor-specific performance results P_i. An LLM then generates a semantic profile $\alpha_{\text{des},i}$ for each factor i:

$$\alpha_{\text{des},i} = F^{\text{profile}}_{\text{LLM}}(M_{\text{global}}, P_i). \tag{2}$$

Each profile articulates the factor's underlying mechanism, its suitability across market regimes (e.g., high-volatility environments), and its limitations or failure conditions. These semantic descriptions serve as an instruction manual for the reasoning core in subsequent decision-making.

3.2.2 Asset Pool State Description

To characterize the current investment environment, we construct a market state description. In contrast to the long-term market memory, this representation is generated dynamically for each decision day t. Using the daily atomic units, $(S^{\text{price}}_t, S^{\text{news}}_t)$, we synthesize the market state up to time t:

$$S_t = F^{\text{state}}_{\text{LLM}}(S^{\text{price}}_t, S^{\text{news}}_t). \tag{3}$$

The resulting state captures the prevailing index dynamics, dominant sector themes, and capital flow patterns, providing the situational context for subsequent factor selection decisions.

3.3 The Alpha-R1 Reasoning Model

The Alpha-R1 model serves as the central reasoning agent for factor screening and selection. At each decision time t, a semantic decision context C_t is constructed by combining two components:

$$C_t = \{\alpha_{\text{des},i}\}_{i \in \mathcal{U}} \oplus S_t, \tag{4}$$

where $\oplus$ represents the structured concatenation of semantic descriptions into a unified prompt context. The components are defined as:

- $\alpha_{\text{des},i}$ denotes the semantic factor description for each candidate factor in the pool $\mathcal{U}$ (sourced from Alpha101 (Kakushadze, 2016), see Section 4.1 for details).
- S_t is the contemporaneous semantic market state synthesized via $F^{\text{state}}_{\text{LLM}}$ as defined in Equa-

tion 3, which encapsulates price dynamics and news narratives at time t.

This context is instantiated into a structured prompt (as detailed in Appendix A.3) that explicitly instructs the model to perform a step-by-step analysis before outputting the final factor selection. Based on this high-dimensional semantic context C_t, Alpha-R1 performs inference to output the final selected factor list, denoted as $\mathcal{A}_t$. Theoretically, we interpret this mechanism as a context-conditioned gating system. Unlike traditional linear models with fixed weights, our approach uses the LLM as a dynamic selector that activates factors based on the semantic alignment between factor profiles and the prevailing market state. By conditioning factor activation on richer semantic information and enforcing parsimonious selection, Alpha-R1 achieves more robust adaptation to regime shifts without the instability of purely numerical re-estimation.

3.4 Reinforcement Learning via GRPO with Market Feedback

We optimize the Alpha-R1 reasoning model using reinforcement learning. This design enables learning directly from objective market feedback, adapting the RLHF paradigm to replace subjective human preferences with performance-based financial signals. The reward function combines quantitative portfolio outcomes with assessments of reasoning quality, and training is carried out using Group Relative Policy Optimization (GRPO) to ensure stable and efficient policy updates.

3.4.1 Backbone Model

We adopt Qwen3-8B as the backbone model for its strong reasoning capabilities. This initialization accelerates convergence during reinforcement learning and improves the consistency and structure of generated outputs. In the absence of such a warm start, models are prone to overfitting superficial heuristics, leading to unstable or incoherent reasoning. The pre-trained backbone provides a stable foundation that preserves prior knowledge, allowing reinforcement learning to refine the model's reasoning behavior.

3.4.2 Multi-Component Reward Function with Market Feedback

We design a multi-component reward function that balances market performance with reasoning discipline:

$$r_{\text{final}} = r_{\text{adjusted}} - P_{\text{structural}}, \tag{5}$$

where r_{adjusted} captures market-based performance feedback, and $P_{\text{structural}}$ denotes the structural penalties that regulate action validity and sparsity. The reward components are computed through the following pipeline.

Rule-based Performance Reward Market feedback is obtained via a backtesting procedure based on a linear factor model trained on four years of historical data. We adopt a linear specification for three reasons: it provides a stable and interpretable mapping from factor exposures to expected returns, enables direct attribution of performance to selected factors, and avoids introducing non-stationarity into the reward signal during reinforcement learning by keeping the evaluation model fixed.

1. Linear Model: We use fixed regression coefficients β_i estimated from historical data.

2. For each stock, predicted returns are computed using the selected factor set $\mathcal{A}_t$:

$$R_{\text{pred}} = \beta_0 + \sum_{i \in \mathcal{A}_t} (\beta_i \times V_i), \tag{6}$$

   where V_i denotes the previous-day value of factor i, and unselected factors contribute zero.

3. Portfolio Construction: Stocks are ranked by predicted returns, and the top N are selected to form an equal-weighted portfolio.

4. Base Reward Calculation: Compute the active return over the benchmark over a holding period H, scaled for the reward function:

$$r_{\text{base}} = (R_{\text{port}} - R_{\text{bench}}) \times 100, \tag{7}$$

   where H denotes the holding period (e.g., $H = 5$ days) used for both the portfolio and the benchmark returns.

Quality-Adjusted Reward with LLM-as-Judge Evaluation We incorporate reasoning quality into the reward through LLM-as-judge evaluation, in which an external large language model automatically assesses the model's generated reasoning. A consistency penalty $P_{\text{consistency}}$ is computed as

$$P_{\text{consistency}} = F_{\text{LLM}}^{\text{judge}}(C_t, \mathcal{A}_t, \text{response}), \tag{8}$$

where $F_{\text{LLM}}^{\text{judge}}$ denotes a judge LLM (Claude 3.5 Haiku) that evaluates dimensions such as logical coherence, linguistic fluency, and information redundancy, yielding a penalty score in the range $[0, 10]$. The variable response represents the full textual output generated by the Alpha-R1 reasoning core, which encompasses both the chain-of-thought reasoning process and the final selected factor list $\mathcal{A}_t$. The resulting score is normalized as $P_{\text{norm}} = P_{\text{consistency}}/10$ and applied asymmetrically to adjust the base reward:

$$r_{\text{adjusted}} = r_{\text{base}} - |r_{\text{base}}| \times P_{\text{norm}}. \tag{9}$$

Structural Penalties The structural penalty $P_{\text{structural}}$ enforces output discipline, which qualitatively combines the requirements for parsimony and validity. It encourages the model to select a concise set of factors to avoid over-complexity, while strictly penalizing the generation of unparsable or non-existent factors to ensure the reasoning results are executable within the quantitative backtesting framework. We define the penalty as:

$$P_{\text{structural}} = 5 \cdot \mathbb{I}(\text{response} \notin \Omega_{\text{valid}}), \tag{10}$$

where $\mathbb{I}(\cdot)$ is the indicator function, and Ω_{valid} represents the set of responses that strictly adhere to the required XML output format (as detailed in Appendix A.3) and contain valid factor identifiers from the universe $\mathcal{U}$. This imposes a hard constraint to prevent format hallucinations.

3.4.3 GRPO Optimization with Market-Aligned Objectives

We employ Group Relative Policy Optimization (GRPO) to fine-tune the Alpha-R1 model. GRPO optimizes the policy by estimating advantages from a group of sampled outputs and applying a clipped surrogate objective to ensure stable updates. This approach allows the model to learn from the multi-component reward function, combining market performance and reasoning quality.

For the detailed mathematical formulation of GRPO, including the normalized advantage estimation and the complete objective function derivation, please refer to Appendix B.

3.5 Portfolio Construction and Execution

To ensure market realism, we implement a slot rotation strategy with Volume-Weighted Average Price (VWAP) execution and transaction costs. Detailed mathematical formulations are provided in Appendix C.

4 Experiments

This section presents a comprehensive empirical evaluation of Alpha-R1. We benchmark Alpha-R1 against a range of traditional quantitative strategies and state-of-the-art large language models (LLMs) to assess its effectiveness in dynamic market environments.

4.1 Experimental Setup

We evaluate Alpha-R1 on four asset pools: S&P 500 and CSI 300 (primary), and Russell 2000 and CSI 1000 (generalization). The dynamic factor zoo is constructed from the Alpha101 library (Kakushadze, 2016). To ensure robust evaluation, we enforce strict temporal separation between factor coefficient estimation (2020–2023), model training (2024), and out-of-sample testing (2025).

We compare against Traditional Quant Strategies (PCA, XGBoost, LightGBM, A2C, PPO) and Reasoning LLMs (Gemini 2.5 Pro, Claude 3.7 Sonnet, DeepSeek-R1, Qwen3-8B). All strategies follow a consistent execution protocol ($H = 5$, $TopN = 10$, VWAP). Detailed configurations for datasets, factor sampling, and model hyperparameters are provided in Appendix E.

Performance is evaluated using cumulative return (CR), annualized return (AR), Sharpe ratio (SR), and maximum drawdown (MDD) (detailed definitions are provided in Appendix D). To reduce the impact of random initialization, all reported backtesting results are averaged over five independent runs.

4.2 Performance Evaluation

We first evaluate the capacity of Alpha-R1 to generate active returns against the benchmark. Table 1 details the quantitative metrics, while Figure 3 visualizes the wealth accumulation trajectories for the main experiments.

4.2.1 Main Results Analysis

As shown in Table 1 and Figure 3, Alpha-R1 consistently outperforms all baselines. Tree-based models (XGBoost, LightGBM) perform poorly, often yielding negative active returns on S&P 500, which highlights the risk of model misspecification in non-stationary markets. Reinforcement learning baselines (A2C, PPO) also struggle; while PPO shows positive returns on CSI 300, it significantly lags behind our model in Sharpe Ratio. In contrast, Alpha-R1 achieves superior risk-adjusted performance with Sharpe Ratios of 1.36 (S&P 500) and

Table 1: Main Experiment Results. Performance comparison of Alpha-R1 against baselines across two asset pools (S&P 500 and CSI 300) (Testing Period: 2025.01.01 – 2025.06.30). Results are reported as the average of 5 independent runs. CR: Cumulative Return, AR: Annualized Return, SR: Sharpe Ratio, MDD: Maximum Drawdown. The best results are highlighted in bold.

Type	Method	Asset Pool S&P 500				Asset Pool CSI 300			
		CR (%)	AR (%)	SR	MDD (%)	CR (%)	AR (%)	SR	MDD (%)
Non-LLM	Buy & Hold	6.44	13.75	0.38	18.75	3.03	6.64	0.33	10.49
	PCA	-1.92	-3.93	-0.39	**14.59**	-5.51	-11.48	-0.81	12.31
	XGBoost	-4.15	-8.37	-0.59	19.21	-9.77	-19.86	-1.42	14.10
	LightGBM	-3.96	-8.00	-0.53	18.89	0.47	1.02	-0.03	11.18
	A2C	2.42	5.05	0.03	15.14	4.62	10.22	0.50	10.02
	PPO	-3.51	-7.11	-0.43	17.00	5.62	12.50	0.68	9.07
LLM	Gemini 2.5 Pro Thinking	0.40	0.82	-0.17	17.20	-11.81	-23.71	-1.63	16.77
	Claude 3.7 Sonnet Thinking	-2.38	-4.89	-0.39	19.77	-9.05	-18.48	-1.30	13.99
	DeepSeek-R1	-0.71	-1.46	-0.27	17.37	-8.55	-17.52	-1.18	16.16
	Qwen3-8B	-2.93	-5.97	-0.48	16.25	-7.63	-15.72	-0.86	18.59
	Alpha-R1 (Ours)	**18.93**	**43.07**	**1.36**	18.41	**14.02**	**32.65**	**1.87**	**6.55**

1.87 (CSI 300) while maintaining competitive maximum drawdowns (lowest on CSI 300), validating the efficacy of our semantic gating mechanism.

4.2.2 Zero-Shot Generalization

Table 2 and Figure 8 present the results for zero-shot generalization (detailed curves and analysis are provided in Appendix F).

On the Russell 2000 universe, Alpha-R1 achieves an exceptional Cumulative Return of 47.32% and a Sharpe Ratio of 3.45, far surpassing the Buy & Hold return of -1.85%. This indicates that the semantic patterns learned from the S&P 500 (large-cap) transfer effectively to small-cap US stocks, potentially capturing fundamental economic drivers that transcend market cap size.

Table 2: Zero-Shot Generalization Results. Performance on out-of-domain asset pools (Russell 2000 and CSI 1000). (Testing Period: 2025.01.01 – 2025.06.30).

Method	Asset Pool Russell 2000			Asset Pool CSI 1000		
	CR (%)	SR	MDD (%)	CR (%)	SR	MDD (%)
Buy & Hold	-1.85	-0.31	23.79	9.64	0.79	16.87
PCA	-31.85	-2.10	35.36	-9.26	-0.82	21.05
XGBoost	-18.72	-1.41	27.31	-13.04	-1.21	22.33
LightGBM	-33.88	-2.04	37.72	-12.03	-1.13	20.72
A2C	20.15	1.76	**13.18**	-10.24	-0.99	22.50
PPO	-42.35	-2.49	43.35	-4.60	-0.45	19.59
Gemini 2.5 Pro Thinking	-44.20	-2.07	45.08	-17.61	-1.37	28.27
Claude 3.7 Sonnet Thinking	-40.37	-1.78	43.91	-20.79	-0.67	26.33
DeepSeek-R1	-45.25	-2.19	46.80	-20.30	-1.52	30.58
Qwen3-8B	-47.71	-2.16	48.73	-5.99	-0.48	22.74
Alpha-R1 (Ours)	**47.32**	**3.45**	16.38	**23.99**	**2.34**	**14.61**

Similarly, on the CSI 1000, Alpha-R1 maintains a solid performance with a Sharpe Ratio of 2.34, outperforming the negative or low returns of most baselines. This confirms the advantage of delegating non-stationarity adaptation to the LLM. By conditioning factor activation on richer state information and enforcing parsimonious selection, our approach reduces misspecification and enables robust zero-shot transfer.

4.3 Ablation Study

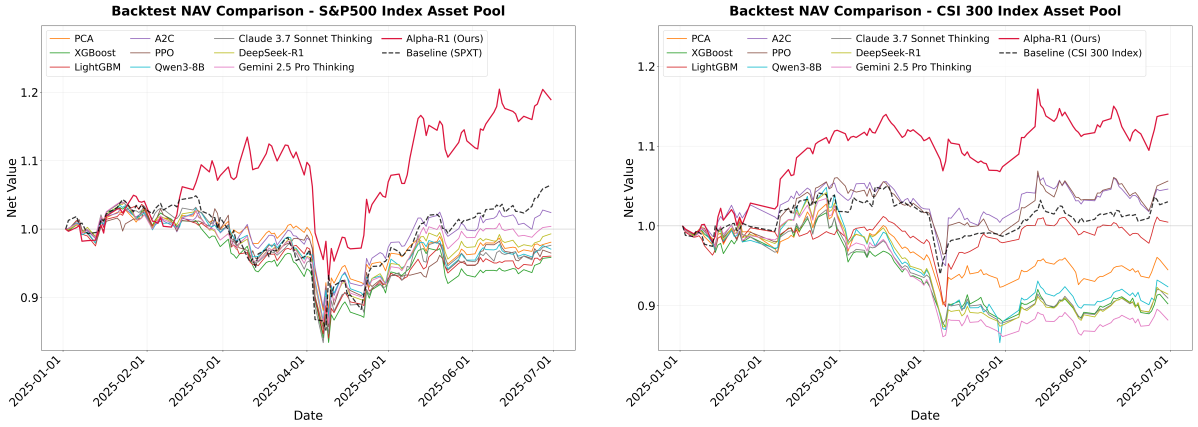
To dissect the contributions of individual components, we conduct a series of ablation experiments on both CSI 300 and S&P 500 asset pools. All ablation variants are trained from scratch using the modified framework to ensure a fair comparison.

4.3.1 Ablation Variants

We evaluate four variants to dissect the contribution of each component: (1) w/o Market Price, which removes market price trends from S_t; (2) w/o News, which excludes market news information from the state space; (3) w/o Factor Semantics, where natural language factor descriptions are replaced with their raw mathematical formulas; and (4) w/o RL Optimization, which uses the base backbone model (Qwen3-8B) without the reinforcement learning alignment pipeline.

4.3.2 Results and Analysis

Table 3 quantifies component contributions. RL Alignment is critical: the unaligned base model fails to generate alpha, proving that general reasoning is insufficient for financial tasks without market-aligned optimization. Factor Semantics also prove vital; replacing descriptions with raw formulas degrades the Sharpe Ratio across both pools, confirming Alpha-R1's reliance on economic rationales. Finally, removing Price or News signals results in lower risk-adjusted returns, highlighting the syn-



(a) Net Value - Asset Pool S&P 500

(b) Net Value - Asset Pool CSI 300

Figure 3: Main Experiment Cumulative Net Value Curves. Comparison of wealth accumulation trajectories on the 2025 testing set. (a) On the S&P 500, Alpha-R1 achieves superior returns compared to baselines. (b) On the CSI 300, Alpha-R1 consistently outperforms traditional strategies and LLMs with lower drawdown.

ergy of our multimodal reasoning core.

Table 3: Ablation Study Results. Performance contribution of different components on CSI 300 and S&P 500 (Testing Period: 2025.01.01 – 2025.06.30).

Method	S&P 500			CSI 300		
	CR (%)	SR	MDD (%)	CR (%)	SR	MDD (%)
Alpha-R1 (Full)	**18.93**	**1.36**	18.41	**14.02**	**1.87**	**6.55**
Buy & Hold	6.44	0.38	18.75	3.03	0.33	10.49
w/o Market Price	11.34	0.81	**15.21**	9.96	1.30	8.55
w/o News	12.40	0.86	17.85	10.18	1.20	9.87
w/o Factor Semantics	10.69	0.70	17.83	6.04	0.72	6.69
w/o RL Optimization	-2.93	-0.48	16.25	-7.63	-0.86	18.59

4.4 Semantic vs. Heuristic Gating Strategies

To validate the superiority of our semantic gating mechanism, we compare it against traditional heuristic gating strategies, including Lasso (a classic sparse linear model selecting factors via L_1 regularization) and IC Momentum (a heuristic selecting the top 10 factors based on their recent average IC over a 20-day window).

Table 4: Gating Strategy Comparison. Performance of Alpha-R1 (Semantic Gating) versus heuristic gating methods (Lasso and IC Momentum) on the CSI 300 and S&P 500 testing sets.

Method	S&P 500			CSI 300		
	CR (%)	SR	MDD (%)	CR (%)	SR	MDD (%)
Alpha-R1	**18.93**	**1.36**	18.41	**14.02**	**1.87**	**6.55**
Lasso	-1.86	-0.35	17.36	1.58	0.13	10.67
IC Momentum	9.69	0.74	**13.99**	-3.64	-0.52	14.16

As shown in Table 4, Alpha-R1 significantly outperforms heuristic baselines on both asset pools. On CSI 300, Lasso achieves a small positive return (1.58%) and IC Momentum performs poorly (-3.64%), while Alpha-R1 reaches 14.02% CR. On S&P 500, while IC Momentum captures some trend (9.69% CR), Alpha-R1 still achieves nearly double the return (18.93%) and a significantly higher Sharpe Ratio (1.36 vs 0.74), demonstrating superior robustness to regime shifts.

4.5 Parameter Sensitivity Analysis

Sensitivity analysis (Appendix G) confirms that Alpha-R1 maintains robust performance across varying holding periods and portfolio sizes. The broad stability region visualized in the heatmaps (Figure 9) indicates that the model's performance is a result of robust semantic reasoning rather than an artifact of overfitting.

5 Conclusion

This paper introduces Alpha-R1, a semantics-driven framework that replaces static factor mining with context-aware economic reasoning. By integrating long-term market memory with real-time news narratives, Alpha-R1 bridges unstructured information and quantitative decision-making in a unified manner. Methodologically, we utilize a market-aligned reinforcement learning approach via GRPO, where training is guided by objective financial outcomes rather than subjective feedback. Crucially, our findings validate that grounding factor screening in semantic reasoning and market feedback provides a viable path to address non-stationarity and alpha decay in financial markets. Empirical results demonstrate that our model consistently outperforms traditional baselines and generic LLMs, exhibiting robust zero-shot generalization in high-volatility environments.

6 Limitations

While Alpha-R1 demonstrates superior performance, we acknowledge several limitations. The inference latency of Large Language Models restricts the framework to low-frequency strategies, making it unsuitable for high-frequency trading. The deployment also entails significant computational overhead due to the intensive reinforcement learning alignment and periodic fine-tuning required for non-stationary markets. Furthermore, the reliance on textual inputs introduces risks related to information quality and potential model hallucinations. These factors necessitate the continued use of traditional risk management layers to ensure safety in extreme market conditions.

References

Bokai Cao, Saizhuo Wang, Xinyi Lin, Xiaojun Wu, Haohan Zhang, Lionel M. Ni, and Jian Guo. 2025. From deep learning to LLMs: A survey of AI in quantitative investment. *Preprint*, arXiv:2503.21422.

Lang Cao. 2025. Chain-of-alpha: Unleashing the power of large language models for alpha mining in quantitative trading. ArXiv:2508.06312.

Mark M. Carhart. 1997. On persistence in mutual fund performance. *The Journal of Finance*, 52(1):57–82.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomek Korbak, David Lindner, Pedro Freire, Tony Tong Wang, Samuel Marks, Charbel-Raphael Segerie, Micah Carroll, Andi Peng, Phillip J.K. Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, and 13 others. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*. Survey Certification, Featured Certification.

Chao Cheng, Bin Chen, Zhe Xiao, and 1 others. 2024. Quantum finance and fuzzy reinforcement learning-based multi-agent trading system. *International Journal of Fuzzy Systems*, 26:2224–2245.

Hongjun Ding, Binqi Chen, Jinsheng Huang, Taian Guo, Zhengyang Mao, Guoyi Shao, Lutong Zou, Luchen Liu, and Ming Zhang. 2025. AlphaEval: A comprehensive and efficient evaluation framework for formula Alpha mining. *Preprint*, arXiv:2508.13174.

Eugene F. Fama and Kenneth R. French. 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56.

Gang Feng, Stefano Giglio, and Dacheng Xiu. 2020. Taming the factor zoo: A test of new factors. *The Journal of Finance*, 75(3):1327–1370.

Joachim Freyberger, Alejandro J. Salgado, and Andriy Shkilko. 2020. Dissecting characteristics nonparametrically. *The Review of Financial Studies*, 33(5):2326–2377.

Shihao Gu, Bryan Kelly, and Dacheng Xiu. 2020. Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638.

Tian Guo and Emmanuel Hauptmann. 2024. Fine-tuning large language models for stock return prediction using newsflow. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1028–1045, Miami, Florida, US. Association for Computational Linguistics.

Jinghai He, Cheng Hua, Chunyang Zhou, and Zeyu Zheng. 2025. Reinforcement-learning portfolio allocation with dynamic embedding of market information. *arXiv preprint arXiv:2501.17992*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*.

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 *International Conference on Learning Representations*.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, Singapore. Association for Computational Linguistics.

Zura Kakushadze. 2016. 101 formulaic alphas. *Wilmott*, 2016(84):72–81.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2025. A survey of reinforcement learning from human feedback. *Transactions on Machine Learning Research*. Survey Certification.

Kelvin J.L. Koa, Yunshan Ma, Ritchie Ng, and Tat-Seng Chua. 2024. Learning to generate explainable stock predictions using self-reflective large language models. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 4304–4315. ACM.

Zhizhuo Kou, Holam Yu, Junyu Luo, Jingshu Peng, Xujia Li, Chengzhong Liu, Juntao Dai, Lei Chen, Sirui Han, and Yike Guo. 2025. Automate strategy finding with LLM in quant investment. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18517–18533, Suzhou, China. Association for Computational Linguistics.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. RewardBench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797, Albuquerque, New Mexico. Association for Computational Linguistics.

Haohang Li, Yangyang Yu, Zhi Chen, Yuechen Jiang, Yang Li, Denghui Zhang, Rong Liu, Jordan W. Suchow, and Khaldoun Khashanah. 2024. Finmem: A performance-enhanced LLM trading agent with layered memory and character design. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

Yang Li, Yangyang Yu, Haohang Li, Zhi Chen, and Khaldoun Khashanah. 2023. Tradinggpt: Multi-agent system with layered memory and distinct characters for enhanced financial trading performance. *Preprint*, arXiv:2309.03736.

Yuante Li, Xu Yang, Xiao Yang, Xisen Wang, Weiqing Liu, and Jiang Bian. 2025. R&d-agent-quant: A multi-agent framework for data-centric factors and model joint optimization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Xiao-Yang Liu, Guoxuan Wang, Hongyang Yang, and Daochen Zha. 2023. FinGPT: Democratizing internet-scale data for financial large language models. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Zhaowei Liu, Xin Guo, Fangqi Lou, Lingfeng Zeng, Jinyi Niu, Zixuan Wang, Jiajie Xu, Weige Cai, Ziwei Yang, Xueqian Zhao, Chao Li, Sheng Xu, Dezhi Chen, Yun Chen, Zuo Bai, and Liwen Zhang. 2025. Fin-r1: A large language model for financial reasoning through reinforcement learning. *Preprint*, arXiv:2503.16252.

Alejandro Lopez-Lira and Yuehua Tang. 2023. Can chatgpt forecast stock price movements? return predictability and large language models. ArXiv:2304.07619.

Dakuan Lu, Hengkui Wu, Jiaqing Liang, Yipei Xu, Qianyu He, Yipeng Geng, Mengkun Han, Yingsi Xin, and Yanghua Xiao. 2023. Bbt-fin: Comprehensive construction of chinese financial domain pre-trained language model, corpus and benchmark. *Preprint*, arXiv:2302.09432.

Jifang Mai, Shaohua Zhang, Haiqing Zhao, and Lijun Pan. 2024. Factor investment or feature selection analysis? *Mathematics*, 13(1):9.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. SimPO: Simple preference optimization with a reference-free reward. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Lingfei Qian, Weipeng Zhou, Yan Wang, Xueqing Peng, Han Yi, Yilun Zhao, Jimin Huang, Qianqian Xie, and Jian yun Nie. 2025. Fino1: On the transferability of reasoning-enhanced llms and reinforcement learning to finance. *Preprint*, arXiv:2502.08127.

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

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

William F. Sharpe. 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, 19(3):425–442.

Hao Shi, Weili Song, Xinting Zhang, Jiahe Shi, Cuicui Luo, Xiang Ao, Hamid Arian, and Luis Angel Seco. 2025. Alphaforge: a framework to mine and dynamically combine formulaic alpha factors. AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Ziyi Tang, Zechuan Chen, Jiarui Yang, Jiayao Mai, Yongsen Zheng, Keze Wang, Jinrui Chen, and Liang Lin. 2025. Alphaagent: Llm-driven alpha mining with regularized exploration to counteract alpha decay. KDD '25, page 2813–2822, New York, NY, USA. Association for Computing Machinery.

Hariom Tatsat and Ariye Shater. 2025. Beyond the black box: Interpretability of llms in finance. *Preprint*, arXiv:2505.24650.

Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Meiyun Wang, Kiyoshi Izumi, and Hiroki Sakaji. 2024. LLMFactor: Extracting profitable factors through prompts for explainable stock movement prediction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3120–3131, Bangkok, Thailand. Association for Computational Linguistics.

Saizhuo Wang, Hang Yuan, Leon Zhou, Lionel Ni, Heung-Yeung Shum, and Jian Guo. 2025. Alpha-GPT: Human-AI interactive alpha mining for quantitative investment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 196–206, Suzhou, China. Association for Computational Linguistics.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *Preprint*, arXiv:2303.17564.

Yijia Xiao, Edward Sun, Tong Chen, Fang Wu, Di Luo, and Wei Wang. 2025a. Trading-r1: Financial trading with llm reasoning via reinforcement learning. *Preprint*, arXiv:2509.11420.

Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. 2025b. Tradingagents: Multi-agents LLM financial trading framework. In *The First MARW: Multi-Agent AI in the Real World Workshop at AAAI 2025*.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: a large language model, instruction data and evaluation benchmark for finance. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Hang Yuan, Saizhuo Wang, and Jian Guo. 2024. Alpha-gpt 2.0: Human-in-the-loop ai for quantitative investment. ArXiv:2402.09746.

Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. 2023a. Instruct-fingpt: Financial sentiment analysis by instruction tuning of general-purpose large language models. *Preprint*, arXiv:2306.12659.

Wentao Zhang, Lingxuan Zhao, Haochong Xia, Shuo Sun, Jiaze Sun, Molei Qin, Xinyi Li, Yuqing Zhao, Yilei Zhao, Xinyu Cai, Longtao Zheng, Xinrun Wang, and Bo An. 2024. A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 4314–4325, New York, NY, USA. Association for Computing Machinery.

Xuanyu Zhang, Qing Yang, and Dongliang Xu. 2023b. Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters. *Preprint*, arXiv:2305.12002.

A Detailed Data Flow

This section provides a comprehensive breakdown of the two primary data streams utilized in Alpha-R1: Price Market Data and News Data. We include examples from both the Chinese A-share market and the US stock market to demonstrate the universality of our approach. The chapter is organized into three main components: Section A.1 details the technical construction methodologies for each data pipeline, Section A.2 presents annotated examples of the actual generated data outputs, and Section A.3 demonstrates the complete prompt structure used for training data generation.

A.1 Construction Methodologies

A.1.1 Price Market Data Flow

The Price Market Data module transforms high-frequency numerical trading data into semantic descriptions using a multimodal large model approach.

- **Visual Encoding:** We retrieve raw OHLC (Open, High, Low, Close) and volume data via the Tushare API (for Chinese markets) and the Massive API (for US markets). We generate two distinct visualizations: (1) A 90-Day K-line Chart capturing medium-term trends, Moving Averages (MA5, MA10, MA20), and volume shifts; and (2) An Intraday Chart capturing minute-level price fluctuations and real-time volume dynamics for the target date.
- **Multimodal Prompting:** These charts are converted to Base64 format and fed into a multimodal large model (Claude Sonnet 3.7) alongside a textual summary of key statistics (e.g., amplitude, turnover rate, price change).
- **Semantic Extraction:** The multimodal large model acts as a technical analyst, outputting a structured description of support/resistance levels, trend signals, and divergence patterns, effectively filtering out noise from the raw numerical series.

A.1.2 News Data Flow

The News Data module processes unstructured textual streams to extract market-relevant narratives, converting raw noise into interpretable signals.

- **Data Aggregation:** Daily financial news reports are collected from major financial news aggregators for the target trading window: Sina Finance for Chinese markets and the Massive API (for US markets), ensuring coverage of macro policies and industry events.
- **LLM Summarization:** A large language model processes the raw text stream to filter out irrelevant information (e.g., advertisements, duplicate reports). It condenses the information into a structured summary focusing on three dimensions: *Macroeconomic Policy* (Central Bank actions, GDP data), *Industry Dynamics* (Sector-specific regulations), and *Significant Corporate Events* (M&A, Earnings).

A.2 Detailed Data Samples

We present representative samples of the generated semantic data below based on real backtesting data from April 8, 2024. These text boxes represent the actual inputs processed by the Alpha-R1 agent (State $S_{20240409}$).

A.3 Detailed Prompt Sample

This section presents the complete prompt structure used for alpha factor selection. The prompt integrates the market environment information (Section A.1) and factor semantic descriptions into a standardized format. It explicitly instructs the model to perform a step-by-step analysis before outputting the final factor selection, thereby enforcing a structured reasoning process.

B GRPO Mathematical Formulation

In this section, we provide the detailed mathematical formulation of the Group Relative Policy Optimization (GRPO) algorithm used to fine-tune the Alpha-R1 model.

B.1 Advantage Estimation

We optimize the Alpha-R1 reasoning model using reinforcement learning with market feedback. The normalized advantage estimate is computed as:

$$\hat{A}_i = \frac{r_i - \text{mean}(r)}{\text{std}(r)}, \tag{11}$$

$$\rho_t^{(i)}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}, \tag{12}$$

where $\rho_t^{(i)}(\theta)$ denotes the probability ratio between the current and old policies.

[Sample 3] US Price Market Description ($S^{\text{price, US}}_{20240409}$)

Market Overview: The US market experienced a broad, high-volume decline. The growth-oriented Nasdaq-100 ETF (QQQ) underperformed the broader S&P 500 ETF (SPY), exhibiting panic selling characteristics with a unilateral downward trend. Short sellers dominated the session as trading volume significantly exceeded typical levels.
Intraday Characteristics (2024-04-08): (1) Morning (9:00–12:00): Indices opened higher but quickly retraced. QQQ hit an intraday high of $443.14 and SPY $524.98 before entering a volatile decline. (2) Mid-day (12:00–14:00): A rapid sell-off occurred post-noon. QQQ dropped from $435.75 to $426.38 between 12:00-13:00. (3) Late-session (14:00–16:00): Panic selling intensified. The final hour saw peak volume, with QQQ closing at $416.24 (-5.00%) and SPY at $496.74 (-4.81%), both near daily lows.

Figure 4: Semantic description generated from US price market data processed by a large language model, highlighting panic selling and sector divergence.

[Sample 2] News Data Summary ($S^{\text{news}}_{20240409}$)

Corporate Violations & Penalties: (1) Aulion Electronics: Received CSRC penalty for misleading statements regarding the performance of key personnel in its Perovskite project. The company was fined 3 million RMB, and key executives were individually fined. (2) Hengxing Tech: Chairman and GM were fined for short-swing trading involving family members.
Restructuring & Distress: (1) Shimao Group: Received a liquidation petition filed by CCB (Asia) at the Hong Kong High Court involving $\sim$1.58 billion HKD. Stock fell 18.68%. (2) *ST Bugao: Announced industrial investors (Baitu Group, etc.) for restructuring, triggering a stock rebound despite a daily drop.
Industry Dynamics: (1) Gaming: NPPA approved 14 imported games (including titles from Tencent and Perfect World). The market expects stable approval cycles to drive sector growth. (2) Liquor: Rumors of Feitian Moutai wholesale prices dropping below 2600 RMB caused volatility; the sector remains weak despite stable actual terminal prices.

Figure 5: Summarized financial news narratives processed by a large language model, highlighting market-moving events up to 30 minutes before market open on 2024-04-09 (providing context for decision-making on 2024-04-09).

B.2 Objective Function

The GRPO objective function is defined as:

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{q,\{o_i\}}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\Big(\rho_t^{(i)}(\theta)\hat{A}_i, \text{clip}\Big(\rho_t^{(i)}(\theta), 1-\epsilon, 1+\epsilon\Big)\hat{A}_i\Big)\Bigg] - \beta\mathbb{E}_q\left[D_{\text{KL}}\left(\pi_\theta(\cdot \mid q)\|\pi_{\text{ref}}(\cdot \mid q)\right)\right] \tag{13}$$

where:

- q represents the input context, including market state and factor descriptions;
- o_i represents the i-th response in the group of size G;
- π_θ, $\pi_{\theta_{\text{old}}}$, and π_{ref} denote the current, sampling, and reference policies, respectively;
- β controls the KL-divergence regularization strength, and ϵ is the clipping parameter.

The objective function consists of two primary components: (1) a clipped surrogate objective that encourages policy updates towards higher advantage, and (2) a KL-divergence regularization term preventing deviation from the reference model. This ensures stable training while optimizing for the market-aligned reward defined in Section 3.

C Portfolio Construction Details

To bridge the gap between theoretical factor performance and realizable investment returns, a robust portfolio construction and execution framework is essential. Theoretical backtests often ignore market frictions such as transaction costs, liquidity limitations, and price impact, leading to overstated performance metrics that fail to materialize in live trading. To address these discrepancies, we implement a practical execution mechanism that accounts for liquidity constraints and transaction costs. Given the factor set $\mathcal{A}_t$ selected by Alpha-R1, portfolio positions are constructed using a slot rotation strategy and executed via volume-weighted average price (VWAP)–based trading.

C.1 Slot Rotation Mechanism

To mitigate the high turnover costs typically associated with daily rebalancing, we divide the total capital into H independent sub-portfolios (referred to as slots), where H corresponds to the holding period (e.g., $H = 5$ days). On any given trading day t, only the specific slot indexed by $k = t \pmod H$ undergoes rebalancing. Specifically, this involves liquidating the existing positions in slot k and opening new equal-weighted positions in the top N stocks ranked by the predicted returns from

[Sample 4] US News Data Summary ($S^{\text{news, US}}_{20240409}$)

Corporate Regulation: QNRX and TPST were halted pending news releases on the evening of April 8 (ET).
Industry Dynamics: (1) Consumer Staples: Bank of America noted the sector's historical defensive nature during recessions but warned that high prices and weak volume growth could undermine resilience. (2) Analyst Ratings: Scotiabank adjusted targets for WM (raised to $260) and others. Wells Fargo lowered targets for several financials/energy stocks including SHEL (to $83) and SCHW (to $87).
Macroeconomic Policy: Tariff concerns sparked a sell-off in US equities and bonds. The 10-year Treasury yield rebound drove 30-year fixed mortgage rates up to 6.82%, the largest single-day jump of the year.

Figure 6: Summarized US financial news narratives processed by a large language model, highlighting market-moving events up to 30 minutes before market open on 2024-04-09 (providing context for decision-making on 2024-04-09).

[Sample 5] Alpha Factor Selection Prompt Structure (Training Data Format)

System Role Definition:
You are a senior quantitative investment expert, skilled in selecting the most suitable alpha factor combinations based on market environment. You need to analyze current market conditions and the characteristics of each factor to provide scientific and reasonable factor selection recommendations.
User Task Instruction:
Based on the following information, select the most suitable factor combinations for 2024-07-01's trading day for 5-day short-term strategy stock selection (buy at market open, sell at market close after 5 trading days).
Target Date: 2024-07-01
Market Environment Information:
• Previous Trading Day Closing Data (2024-06-28): [Market price data content...]
• Previous Trading Day Market Analysis (2024-06-28): [Technical analysis content...]
• Current Day Pre-Market News (2024-07-01): [Financial news content...]
Available Factor Descriptions:
• alpha001: [Detailed factor description including calculation method, historical performance...]
• alpha002: [Detailed factor description...]
• [Additional factors with complete descriptions...]
Analysis Framework:
1. Analyze each factor's nature and characteristics
2. Evaluate factors' expected performance in current market
3. Make final selection (maximum 10 factors)
Output Requirements:
• Provide detailed analytical reasoning first
• Output XML-tagged factor list: <alpha_list><alpha001>...</alpha_list>
• Maximum 10 factors allowed
• Skip selection if no factors expected to yield positive returns
Expected Response Format:
Factor analysis reasoning: [Detailed explanation of selection logic...]
The most suitable factor selection for the current market is: <alpha_list><alpha001><alpha003><alpha007></alpha_list>

Figure 7: Complete prompt structure for alpha factor selection training. This demonstrates the standardized format used to generate training data for Alpha-R1, incorporating market data and factor descriptions.

the current factor set $\mathcal{A}_t$:

$$P_{t,k} = \text{Top}_N(\text{Rank}(\text{LinearModel}(\mathcal{A}_t))), \quad (14)$$

where $P_{t,k}$ represents the updated holdings of the k-th slot. The remaining $H-1$ slots remain passive. This approach effectively smooths out the equity curve and reduces the average daily turnover rate to $1/H$, allowing for broader market coverage without incurring excessive friction costs.

C.2 VWAP-based Execution and Constraints

Unlike simplified backtesting engines that assume execution at the opening price P_{open}, we employ a volume-weighted average price (VWAP) model to better approximate realized trading costs. For a selected stock s, the execution price $\hat{P}_{s,t}$ is computed using transaction data from the first 30 minutes of the trading session (09:31–10:00):

$$\hat{P}_{s,t} = \frac{\sum_{i=1}^{30}(Price_{s,t,i} \times Volume_{s,t,i})}{\sum_{i=1}^{30} Volume_{s,t,i}}. \quad (15)$$

The 30-minute execution window (09:31–10:00) corresponds to the period of highest market liquidity and provides a conservative estimate of execution slippage. To ensure market realism, the execution process enforces the following constraints:

- Limit-Move Constraints: Buy orders are rejected if the stock hits the upper price limit

(Limit-Up) during the execution window, and sell orders are deferred if the stock is locked at the lower price limit (Limit-Down).

- IPO Exclusion: Stocks are strictly excluded from trading on their initial listing day to avoid extreme volatility distortions.
- Transaction Costs: A transaction fee of 0.1% (10 bps) is applied to both buy and sell orders to account for commissions and slippage.

The daily portfolio return R_t is computed as the aggregate return across all H slots, providing an overall measure of execution-adjusted strategy performance.

D Performance Evaluation Metrics

To thoroughly assess the financial efficacy of the Alpha-R1 strategy, we employ a standard suite of quantitative metrics, covering profitability, risk management, and predictive signal quality. Let V_t be the portfolio value at time t, and R_t the return in period t.

D.1 Profitability and Risk Metrics

D.1.1 Cumulative Return (CR)

The Cumulative Return (CR) quantifies the total percentage gain or loss over the entire evaluation period T.

$$\mathrm{CR} = \prod_{t=1}^{T}(1 + R_t) - 1.$$

D.1.2 Annualized Return (AR)

The Annualized Return (AR) represents the geometric average amount of money earned by an investment each year over a given time period.

$$\mathrm{AR} = (1 + \mathrm{CR})^{\frac{K}{T}} - 1,$$

where K is the number of trading periods per year (e.g., 252 for daily trading) and T is the total number of trading periods in the evaluation.

D.1.3 Sharpe Ratio (SR)

The Sharpe Ratio (SR) measures risk-adjusted performance by comparing excess returns to volatility. Let $R_e = R_p - R_f$ be the excess return, with mean μ_e and standard deviation σ_e.

$$\mathrm{SR}_{\text{period}} = \frac{\mu_e}{\sigma_e}.$$

The annualized SR is $\mathrm{SR}_{\text{ann}} = \mathrm{SR}_{\text{period}} \cdot \sqrt{K}$, where K is the number of trading periods per year.

D.1.4 Maximum Drawdown (MDD)

Maximum Drawdown (MDD) captures the largest peak-to-trough percentage decline in portfolio value over the sample period, reflecting tail risk.

$$\mathrm{MDD} = \max_{t\in[1,T]} \left(\frac{\max_{\tau\in[1,t]} V_\tau - V_t}{\max_{\tau\in[1,t]} V_\tau} \right).$$

D.2 Predictive Signal Quality Metrics

These metrics evaluate the cross-sectional correlation and temporal consistency between the model's predicted factor scores (f_t) and the actual realized future returns (R_{t+1}).

D.2.1 Information Coefficient (IC)

The Information Coefficient (IC) measures the linear correlation between the predicted factor scores f_t and the vector of actual future returns R_{t+1}. It is typically calculated as the Pearson correlation coefficient:

$$\mathrm{IC}_t = \rho_{\text{Pearson}}(f_t, R_{t+1}).$$

D.2.2 Information Coefficient Information Ratio (ICIR)

The IC Information Ratio (ICIR) assesses the temporal consistency of the predictive signal.

$$\mathrm{ICIR} = \frac{\mathbb{E}[\mathrm{IC}_t]}{\sqrt{\mathrm{Var}(\mathrm{IC}_t)}}.$$

D.2.3 Rank Information Coefficient (Rank IC)

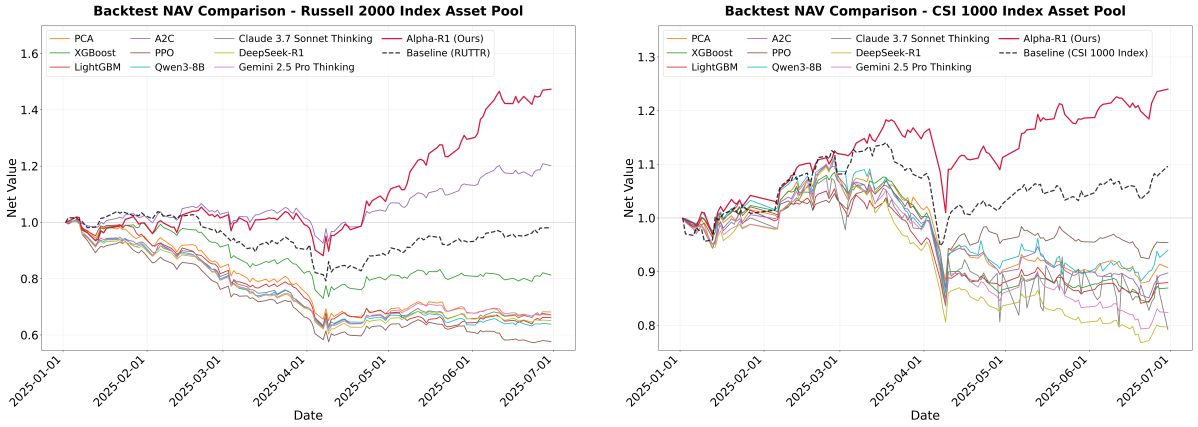
The Rank IC is a robust, non-parametric variant (Spearman's rank correlation coefficient) that measures the monotonic relationship between the ranks of predicted signals and future returns, reducing sensitivity to outliers.

$$\mathrm{Rank\,IC}_t = \rho_{\text{Spearman}}\big(\mathrm{rank}(f_t), \mathrm{rank}(R_{t+1})\big).$$

D.2.4 Rank IC Information Ratio (Rank ICIR)

The Rank ICIR evaluates the temporal stability of the Rank IC.

$$\mathrm{Rank\,ICIR} = \frac{\mathbb{E}[\mathrm{Rank\,IC}_t]}{\sqrt{\mathrm{Var}(\mathrm{Rank\,IC}_t)}}.$$



(a) Net Value - Russell 2000

(b) Net Value - CSI 1000

Figure 8: Zero-Shot Generalization Curves. Comparison of wealth accumulation trajectories on out-of-domain asset pools: (a) Russell 2000 and (b) CSI 1000.

E Model Configurations and Reproducibility

Experimental Setup Details. Experiments are conducted with strict anti-leakage protocols. β_i coefficients for factor synthesis are estimated using historical data from 2020–2023.

- **Training Phase (2024.07.01 – 2024.12.31):** We employ randomized factor augmentation, sampling a subset of 40 factors per episode to foster semantic reasoning capabilities.
- **Testing Phase (2025.01.01 – 2025.06.30):** For out-of-sample evaluation, we fix the candidate pool to the top 40 factors based on their Mean Rank IC during the backtesting period.
- **Parameters:** Risk-free rates are set at 4.5% for the US market and 1.5% for the Chinese market.

Hyperparameter Settings. Alpha-R1 and the baseline LLMs are configured to ensure deterministic and comparable inference. We set the decoding `temperature`=0 and `top_p`=0.7. Alpha-R1 adopts Qwen3-8B as its backbone.

Data Leakage Prevention. To ensure a fair evaluation, all benchmark LLMs have pre-training data cutoffs strictly no later than December 31, 2024. This aligns with our testing phase, which starts on January 1, 2025.

F Zero-Shot Generalization Curves

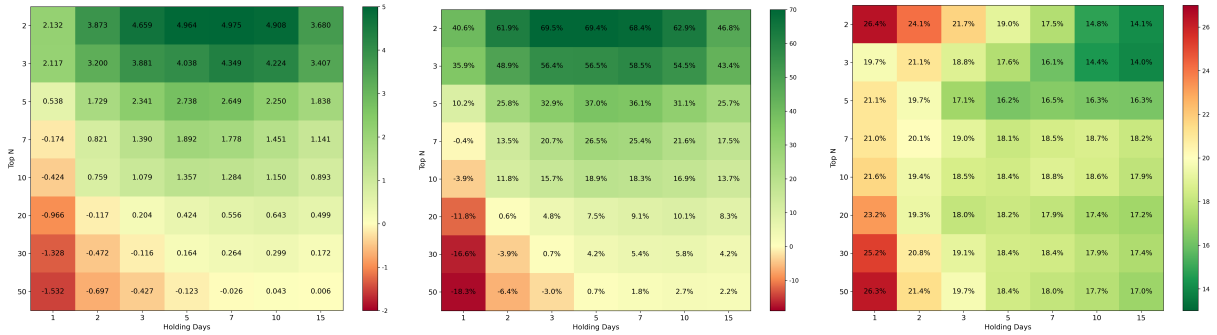
This appendix provides the detailed cumulative net value curves for the zero-shot generalization experiments, supplementing the quantitative results in Section 4. Figure 8 illustrates the wealth accumulation trajectories on the Russell 2000 (US small-cap) and CSI 1000 (China small-cap) asset pools, where the model was applied without any fine-tuning on these specific universes.

On the Russell 2000 (Figure 8a), Alpha-R1 demonstrates a remarkable decoupling from the market index and baseline strategies. While the Buy & Hold benchmark and traditional quantitative models (e.g., PCA, XGBoost) show stagnation or negative returns, our model maintains a consistent upward trend. This visual evidence supports the high Sharpe Ratio reported in the main text, suggesting that the semantic patterns of factor effectiveness learned from large-cap stocks (S&P 500) are robust and transferable to the small-cap domain.

Similarly, for the CSI 1000 (Figure 8b), Alpha-R1 achieves stable growth despite the high volatility and drawdowns observed in the market index. The curves clearly show that while heuristic and tree-based baselines struggle to adapt to the unseen data distribution of the CSI 1000, Alpha-R1 effectively identifies profitable opportunities while managing risk. This confirms the generalization capability of the semantic gating mechanism, which relies on economic reasoning rather than overfitting to specific statistical patterns of the training pool.

G Parameter Sensitivity Analysis

To assess the strategy's resilience to hyperparameter variations, we conduct a comprehensive sensitivity analysis on two critical execution parameters: the holding period (H) and the number of selected stocks per slot ($TopN$). We vary H from 1 to 10



(a) S&P 500: Sharpe Ratio (b) S&P 500: Cumulative Return (c) S&P 500: Max Drawdown

(d) CSI 300: Sharpe Ratio (e) CSI 300: Cumulative Return (f) CSI 300: Max Drawdown

Figure 9: Parameter Sensitivity Analysis on Main Asset Pools. The heatmaps illustrate the impact of varying *TopN* and *H* on S&P 500 and CSI 300. Green regions indicate favorable performance (High Sharpe/Return, Low Drawdown), while Red regions indicate poor performance. The broad Green clusters across both asset pools confirm the strategy's robustness.

days and $TopN$ from 5 to 30 stocks, evaluating the impact on risk-adjusted returns across both S&P 500 and CSI 300 asset pools. The results are visualized via heatmaps in Figure 9.

Impact of Holding Period (H). As illustrated in Figure 9, Alpha-R1 maintains robust performance across a wide range of holding periods. Specifically, on the CSI 300, the Sharpe Ratio remains consistently high (above 1.5) for $H \in [3, 8]$, indicating that the alpha signals generated by our model possess a durable predictive horizon beyond immediate microstructure effects. For the S&P 500, while shorter holding periods ($H \leq 5$) yield slightly higher returns due to faster adaptation to market news, the strategy remains profitable even at $H = 10$, demonstrating that the semantic logic captures fundamental shifts rather than transient noise.

Impact of Portfolio Size ($TopN$). The performance is also remarkably stable with respect to portfolio concentration. Increasing $TopN$ from 5 to 20 results in a smoother equity curve with reduced volatility, as idiosyncratic risk is diversified away. Importantly, the Sharpe Ratio does not degrade significantly as the portfolio expands, suggesting that Alpha-R1 identifies a broad cluster of effective factors rather than relying on a single outlier signal. This "broad green region" in the heatmaps confirms that the model's advantage is not an artifact of overfitting to a specific, narrow parameter configuration, but rather a result of robust semantic reasoning.