

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

Zuoyou Jiang^{*†}

Shanghai Jiao Tong University
Shanghai, China
zuoyou.jiang@sjtu.edu.cn

Ruohan Sun[†]

Shanghai Jiao Tong University
Shanghai, China
ruohan.sun@sjtu.edu.cn

Daxin Jiang

StepFun
Shanghai, China
djiang@stepfun.com

Li Zhao^{*}

StepFun
Shanghai, China
zhaoli@stepfun.com

Zhongjian Li[†]

Shanghai Jiao Tong University
Shanghai, China
zhongjian.li@sjtu.edu.cn

Zuo Bai[‡]

FinStep, StepFun
Shanghai, China
baizuo@{finstep.cn, stepfun.com}

Rui Sun

StepFun
Shanghai, China
sunrui@stepfun.com

Jing Li

StepFun
Shanghai, China
futureli@stepfun.com

Cheng Hua[‡]

Shanghai Jiao Tong University
Shanghai, China
cheng.hua@sjtu.edu.cn

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 show that Alpha-R1 consistently outperforms benchmark strategies and exhibits improved robustness to alpha decay. The full implementation and resources are available at <https://github.com/FinStep-AI/Alpha-R1>.

Keywords

Large Language Models, Reinforcement Learning, Alpha Screening, Quantitative Trading

1 Introduction

The paradigm of factor investing has remained a cornerstone of modern asset management since the seminal work of Fama and French [7]. Theoretical evolution, ranging from the foundational CAPM [37] to multi-factor models [3], coupled with the exponential

growth of market data, has culminated in the high-dimensional phenomenon known as the Factor Zoo [8].

At the same time, advancements in natural language processing (NLP) have empowered the extraction of valuable sentiment signals from extensive unstructured data sources, such as news and financial reports [28]. The field has witnessed a paradigm shift from general-purpose architectures toward domain-specific foundation models. Notable examples, including BloombergGPT [44] and FinGPT [26], have validated the effectiveness of pre-training and fine-tuning on financial datasets, often leveraging efficient adaptation methods like LoRA [16]. These specialized models have set new performance standards [47], demonstrating superior capabilities in tasks such as sentiment analysis and named entity recognition compared to generalist models [29, 49, 51].

However, a key gap remains in unifying these data streams. Traditional numerical indicators and textual signals are often treated as separate modalities, combined through static weighting schemes or heuristic rules rather than within a unified framework that captures their semantic interactions in dynamic decision-making. Recent advances in large language models (LLMs) have substantially expanded the scope of automated reasoning. NLP has shifted from task-specific supervision toward promptable, general-purpose systems that, augmented with chain-of-thought reasoning and reinforcement learning, can address increasingly complex cognitive tasks.

The integration of these systems into factor investing is far from mature. The financial domain is characterized by inherent non-stationarity, noise, and high dimensionality [8, 13], necessitating adaptive reasoning to navigate uncertainty, such as assessing the relative merit of value versus momentum factors under shifting macroeconomic conditions. These demands stand in stark contrast to the deterministic nature of coding or mathematical benchmarks typically employed for LLMs [14]. Although alternative methodologies like sparsity-based machine learning (e.g., Lasso) are available [9], they frequently suffer from poor interpretability and instability during market regime changes. Conversely, general-purpose

^{*}Both authors contributed equally to this research.

[†]Work done during an internship at FinStep and StepFun.

[‡]Corresponding author.

LLMs typically lack alignment with financial principles, making it difficult to produce verifiable justifications for factor selection or transparent decision traces [40].

Current research on LLMs in finance has largely focused on factor mining, namely the discovery of novel signals from textual or multimodal data. This line of work is closely related to open-domain question answering, where the primary goal is information extraction. A growing literature leverages the generative capabilities of LLMs for this purpose. For instance, Cheng et al. [5] showed that GPT-4 can generate factors with high Sharpe ratios supported by economic rationales, while Wang et al. [42] proposed prompt-based extraction methods for profitable factors. To further operationalize factor generation, Cao [2] introduced Chain-of-Alpha, a dual-chain architecture that generates seed factors and iteratively refines them using backtesting feedback. Related frameworks, including Alpha-GPT [43, 48] and R&D-Agent-Quant [24], explore human-in-the-loop and multi-agent systems to bridge hypothesis generation and code implementation. To address alpha decay, Tang et al. [39] proposed regularization based on abstract syntax trees to encourage factor originality. More recently, Kou et al. [21] utilize LLMs to extract trading signals from multimodal data, including financial text and market information.

As factor generation capabilities improve, evaluation has emerged as a key bottleneck. To address this, Ding et al. [6] proposed AlphaEval, a five-dimensional framework that evaluates factors without backtesting. Nevertheless, a fundamental gap remains between broad factor mining and the path-dependent reasoning required for systematic factor screening. Recent efforts such as Trading-R1 [45], Fin-R1 [27], and FinO1 [33] enhance reasoning capabilities in financial LLMs, yet factor screening decisions remain highly context-dependent and time-varying. A factor that performs well in an inflationary regime may become ineffective, or even harmful, during a recession.

To address these structural limitations, we propose Alpha-R1, a dynamic investment framework anchored by a specialized reasoning model trained via reinforcement learning. Distinct from generic LLM applications or purely agentic frameworks [15, 46], 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 to assess the economic relevance of candidate factors and construct portfolios aligned with prevailing market conditions. In addition, Alpha-R1 attributes return sources in a structured manner, enabling transparent explanations of factor selection decisions. This design addresses both the opacity of traditional quantitative models and the static reasoning of existing financial LLMs, advancing regime-aware and interpretable portfolio construction.

Our primary contributions are as follows. First, we develop a practical investment framework that bridges static quantitative models and dynamic market environments. By synthesizing heterogeneous information, including macroeconomic indicators and news narratives, the framework enables regime-aware factor screening and dynamically adjusts portfolio exposure based on the semantic alignment between factor rationales and prevailing market conditions.

Second, we design a specialized reasoning core by adapting the reinforcement learning from human feedback (RLHF) paradigm

to the financial domain. Instead of relying on subjective human preferences, we construct an objective reward signal based on realized market performance, such as volatility-adjusted returns. This design aligns the model’s reasoning process with realistic trading objectives and supports sequential decision-making under uncertainty.

Finally, we conduct extensive backtesting across multiple asset pools beyond standard market indices. The results show that Alpha-R1 consistently outperforms state-of-the-art benchmarks and traditional factor strategies, demonstrating robustness to alpha decay and the ability to deliver explainable, superior risk-adjusted performance across market regimes.

2 Related Work

2.1 LLMs in Quantitative Trading

2.1.1 Financial Foundation Models and Adaptation. Before the advent of reasoning-centric approaches, the adaptation of LLMs to finance primarily focused on domain-specific pre-training and instruction tuning. BloombergGPT [44] established a benchmark by training on a massive mixed corpus of financial and general data. To democratize access, open-source efforts like FinGPT [26] and Instruct-FinGPT [49] utilized efficient fine-tuning techniques to adapt general-purpose models for financial tasks. Additionally, BBT-Fin [29] and XuanYuan 2.0 [51] have explored Chinese financial benchmarks, while PIXIU [47] provided a comprehensive evaluation framework for these instruction-tuned models.

2.1.2 LLMs for Alpha Factor Generation. The integration of LLMs has catalyzed a paradigm shift in quantitative trading, particularly within the domain of alpha factor mining, where models automate the discovery of novel, interpretable predictive signals. Contemporary research has evolved from simple signal generation to sophisticated agent-based frameworks. A prominent research trajectory focuses on mitigating alpha decay through iterative refinement. For instance, AlphaAgent [39] introduces a framework that enforces factor originality and robustness using abstract syntax tree (AST) similarity measures. Complementing this, Chain-of-Alpha [2] proposes a dual-chain architecture consisting of a factor generation chain and a factor optimization chain to iteratively refine candidate factors based on feedback from backtesting. Similarly, AlphaForge [38] adopts a two-stage generative and predictive neural network to enhance the mining process. Beyond fully autonomous systems, recent work has also explored human-AI collaboration [43, 48].

2.1.3 LLM-based Quant Agents. Moving beyond the paradigm of discrete factor mining, the research frontier has expanded towards developing holistic, AI-driven agents capable of orchestrating the entire investment lifecycle. A primary research trajectory focuses on fine-tuning LLMs for direct return forecasting. Guo and Hauptmann [12] demonstrate that adapting models to unstructured data, such as news flow, substantially enhances stock selection capabilities, while also providing critical comparative insights into the efficacy of encoder-only (e.g., DeBERTa) versus decoder-only (e.g., Mistral) architectures.

To address the downstream complexities of execution and risk management, however, the field is increasingly pivoting towards multi-agent orchestration. Kou et al. [21] propose a hierarchical

framework where LLMs function as specialized alpha generators, with distinct agents responsible for dynamic weight optimization. This architectural philosophy that segregates signal generation from risk control parallels the modular design of reinforcement learning systems such as QF-FRL [5] and general multi-agent frameworks like MetaGPT [15]. This trend culminates in comprehensive ecosystems like TradingGPT [25], FinMem [23], and FinAgent [1, 50], which employ multimodal perception [21] and collaborative agent workflows (e.g., TradingAgents [46]) to autonomously navigate the full spectrum of trading activities. Additionally, Koa et al. [20] integrated self-reflection mechanisms to refine predictions, addressing hallucination issues common in these complex agents [17].

2.2 Methods for Screening Alpha Factors

2.2.1 Machine Learning in Regularized and Sparsity-Driven Selection. The most direct machine learning approach to addressing the Factor Zoo is the application of sparsity-inducing regularization, such as Lasso [41]. This method imposes sparsity by adding a penalty term, effectively shrinking the coefficients of irrelevant or redundant factors to zero. Building on this principle, Mai et al. [30] adopt the Lasso method to address severe multicollinearity among feature indicators in the A-share market, finding that it yields superior stability and predictive accuracy compared to traditional unregularized linear methods.

2.2.2 Machine Learning in Tree-Based and Non-Linear Screening. A second pathway leverages tree-based ensemble models. Gu et al. [10] provided a milestone comparison of ML methods, establishing that neural networks and regression trees demonstrate exceptional performance in capturing nonlinear interactions for return prediction. In this approach, factors are typically screened based on feature importance scores. Mai et al. [30] employed tree-based models and deep feedforward Neural Networks to evaluate the efficacy of factors selected by Lasso. Their results confirmed the efficacy of non-linear models in portfolio management, consistently identifying key predictive factors such as momentum and earnings yield across varying market regimes.

2.2.3 Deep Learning in Factor Compression. In contrast to selecting a sparse subset of factors, deep learning techniques offer an alternative by compressing the high-dimensional factor universe into a low-dimensional set of latent representations. This is typically achieved through autoencoders or similar architectures. While effective at dimension reduction, a fundamental limitation lies in the opaque black box nature of these models, which often obscures the economic interpretation of the resulting latent factors. To address this, Mai et al. [30] propose a hybrid CPCA framework that combines clustering with Principal Component Analysis (PCA) to construct lower-dimensional investment factors while maintaining a degree of interpretability often lost in purely deep learning-based approaches.

2.3 Reinforcement Learning for LLMs

Reinforcement learning has established itself as a cornerstone paradigm for aligning LLMs with human intent and augmenting their reasoning faculties [19]. The dominant framework, reinforcement

learning from human feedback (RLHF), was introduced to bolster instruction-following capabilities [32]. This paradigm predominantly relies on proximal policy optimization (PPO) [35], which utilizes a learned value function to estimate advantages and ensure training stability via a clipped surrogate objective. However, the requirement to maintain a value model incurs significant memory and computational overhead. Moreover, traditional RLHF faces challenges such as reward hacking and instability [4, 22]. To mitigate these, alternative preference optimization methods like direct preference optimization (DPO) [34] and simple preference optimization (SimPO) [31] have been proposed to bypass explicit reward modeling.

To circumvent computational bottlenecks in reasoning-intensive tasks, recent methodological advancements have increasingly favored critic-free optimization paradigms. Notably, group relative policy optimization (GRPO) [36] has emerged as a robust alternative. Diverging from PPO, GRPO eschews the need for a separate value network by approximating the baseline via the mean rewards of a group of outputs sampled from the current policy. This group-relative formulation significantly reduces the computational footprint of RL training while preserving optimization stability. DeepSeek-R1 [11] exemplifies the efficacy of this paradigm in fostering self-evolving reasoning chains. This paradigm is now permeating the financial domain; for instance, Trading-R1 [45] utilizes a curriculum of supervised fine-tuning and RL to structure investment theses, while Fin-R1 [27] and FinO1 [33] explore the transferability of reasoning capabilities to financial tasks. This directly motivates our adoption of this critic-free architecture for Alpha-R1, optimizing factor screening logic against objective market feedback.

3 Methodology

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 construct two complementary market descriptors:

- Price Market Description (S_t^{price}): Summarizes 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 M_w is updated recursively using a large language model:

$$M_w = F_{\text{LLM}}(I_w \oplus M_{w-1}). \quad (1)$$

After iterating through the entire backtest period, we obtain the comprehensive backtest period market description (M_{global}). This encapsulates the structural evolution and regime shifts of the

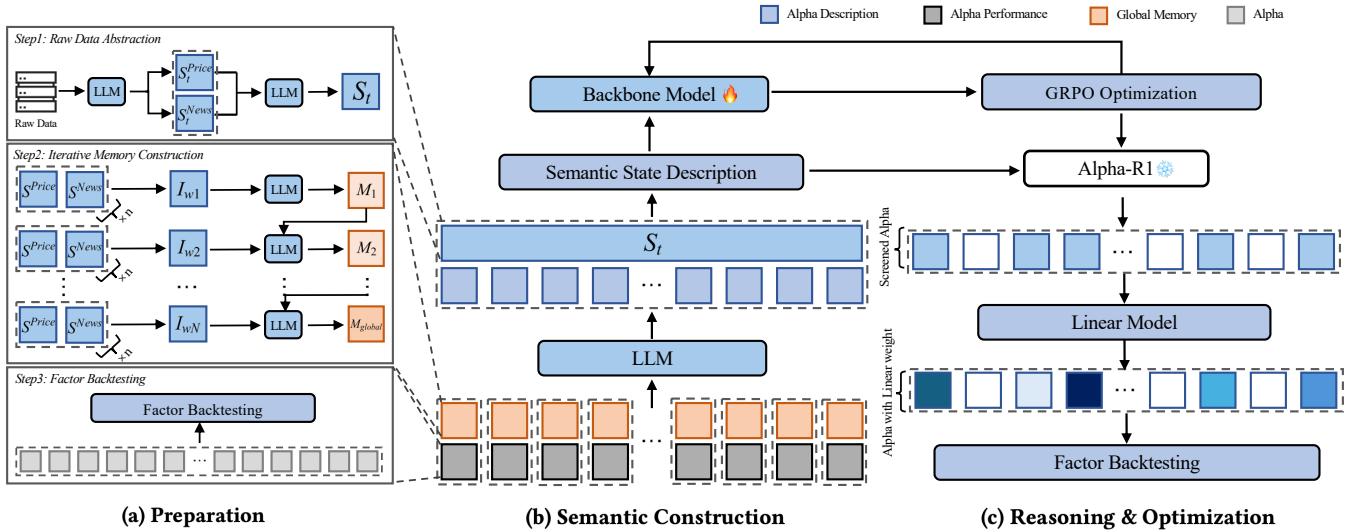


Figure 1: Alpha-R1 Framework Overview. The pipeline follows a sequential logic: (a) Preparation: Abstracting raw technical indicators and financial news into atomic textual units to construct a global historical memory (M_{global}), coupled with systematic factor backtesting; (b) Semantic Construction: 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.

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 backtest period, M_{global} , with the factor-specific backtest results P_i . An LLM then generates a semantic profile $\alpha_{\text{des},i}$ for each factor i :

$$\alpha_{\text{des},i} = F_{\text{LLM}}(M_{\text{global}}, P_i). \quad (2)$$

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 an asset pool 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, $(S_t^{\text{price}}, S_t^{\text{news}})$, we synthesize the instantaneous market state:

$$S_t = F_{\text{LLM}}(S_t^{\text{price}}, S_t^{\text{news}}). \quad (3)$$

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:

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

where:

- $\{\alpha_{\text{des},i}\}$ denotes the set of semantic factor descriptions for the candidate pool \mathcal{U} .
- S_t is the contemporaneous semantic market state synthesized via F_{LLM} as defined in Equation 3, which encapsulates price dynamics and news narratives at time t .

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 . We interpret this mechanism as a context-conditioned gating process where the LLM functions as a network that activates or deactivates factors based on semantic alignment between (i) factor mechanism profiles and failure conditions, and (ii) the current market state summarized from price dynamics and news narratives. This delegation of non-stationarity adaptation to the reasoning core allows the system to navigate regime shifts without the instability of purely numerical re-estimation.

Theoretically, we interpret our framework as a context-conditioned sparse linear model. The fixed linear scorer provides a stable, low-variance mapping from factor exposures to stock ranking. In regime-switching markets, the dominant error source is often model misspecification rather than within-regime estimation. While a purely dynamic linear model must re-estimate coefficients from limited and noisy samples—inducing high variance and overreaction to transient correlations—our approach reduces misspecification and estimation noise. By conditioning factor activation on richer state information and enforcing parsimonious selection, Alpha-R1 achieves more robust out-of-sample performance.

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 and Stability. 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:

$$R_{\text{final}} = R_{\text{adjusted}} - P_{\text{structural}}, \quad (5)$$

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 :

$$\text{Return}_{\text{predicted}} = \beta_0 + \sum_{i \in \mathcal{A}_t} (\beta_i \times V_i), \quad (6)$$

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 excess return over the benchmark over a holding period H , scaled for the reward function:

$$R_{\text{base}} = (\text{Return}_{\text{port}}(\mathcal{A}_t, H) - \text{Return}_{\text{bench}}(H)) \times 100, \quad (7)$$

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-a-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

$$P_{\text{consistency}} = F_{\text{Judge}}(C_t, \mathcal{A}_t, \text{response}), \quad (8)$$

where F_{Judge} denotes a judge LLM (e.g., Claude 3.5 Haiku) that evaluates dimensions such as logical coherence, linguistic fluency, and information redundancy. 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 $P_{\text{norm}} = P_{\text{consistency}} / 10.0$ and applied asymmetrically to adjust the base reward:

$$R_{\text{adjusted}} = \begin{cases} R_{\text{base}} \times (1 - P_{\text{norm}}) & \text{if } R_{\text{base}} > 0 \\ R_{\text{base}} \times (1 + P_{\text{norm}}) & \text{if } R_{\text{base}} \leq 0 \end{cases}. \quad (9)$$

Structural Penalties. We incorporate a comprehensive structural penalty $P_{\text{structural}}$ to enforce output discipline. This term 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.

3.4.3 GRPO Optimization with Market-Aligned Objectives. We employ Group Relative Policy Optimization (GRPO) to fine-tune the Alpha-R1 model. The normalized advantage estimate is computed as:

$$\hat{A}_i = \frac{r_i - \text{mean}(r)}{\text{std}(r)}, \quad \rho_t^{(i)}(\theta) = \frac{\pi_\theta(o_{i,t} | q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} | q, o_{i,<t})}, \quad (10)$$

where $\rho_t^{(i)}(\theta)$ denotes the probability ratio between the current and old policies. The GRPO objective function is defined as:

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{q, \{o_i\}} \left[\frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \min \left(\rho_t^{(i)}(\theta) \hat{A}_i, \right. \right. \\ \left. \left. \text{clip} \left(\rho_t^{(i)}(\theta), 1 - \epsilon, 1 + \epsilon \right) \hat{A}_i \right) \right] \\ - \beta \mathbb{E}_q [D_{\text{KL}}(\pi_\theta(\cdot | q) \| \pi_{\text{ref}}(\cdot | q))], \quad (11)$$

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 , π_{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 Equation 5. Through this approach, Alpha-R1 learns to select factor combinations that not only generate superior excess returns but also exhibit coherent and interpretable reasoning patterns.

3.5 Portfolio Construction and Execution

To translate factor selection into realized market performance, 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.

3.5.1 Slot Rotation Mechanism. To mitigate the high turnover costs typically associated with daily rebalancing, we divide the total capital C 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:

$$P_{t,k} = \text{Rebalance}(P_{t-1,k}, \mathcal{A}_t), \quad (12)$$

where $P_{t,k}$ represents the 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.

3.5.2 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):

$$\hat{P}_{s,t} = \frac{\sum_{i=1}^{30} (\text{Price}_{s,t,i} \times \text{Volume}_{s,t,i})}{\sum_{i=1}^{30} \text{Volume}_{s,t,i}}. \quad (13)$$

This interval 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.

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. Our evaluation is guided by the following research questions:

Q1: Does Alpha-R1 consistently outperform traditional machine learning baselines and reasoning LLMs across distinct asset pools?

Q2: What are the contributions of individual components (e.g., news, price, semantic descriptions) to the model's performance?

Q3: Does the semantic gating mechanism of Alpha-R1 offer tangible advantages over traditional heuristic gating strategies?

Q4: How sensitive is the model's performance to hyperparameter variations?

4.1 Experimental Setup

4.1.1 Dataset and Dynamic Factor Zoo. We construct a dynamic factor zoo by filtering the Alpha101 library [18], retaining 82 computationally feasible factors. To avoid look-ahead bias, the experiments are temporally segmented:

Pre-training Phase (2020.01.01 – 2023.12.31): Used solely to estimate the fixed coefficients β_i for the linear reward model described in Section 3.4.

Training Phase (2024.07.01 – 2024.12.31): Using constituent stocks from the CSI 300, we simulate a high-turnover environment. To encourage semantic reasoning rather than factor memorization, we adopt a randomized factor augmentation scheme: for each trading date, 300 training samples are generated, each with a randomly selected subset of 40 factors drawn from the full pool of 82 as the state input.

Testing Phase (2025.01.01 – 2025.06.30): For strictly out-of-sample evaluation, we construct a fixed candidate set of 40 factors selected based on historical RankIC performance over 2023–2024. Evaluation is conducted on both the CSI 300 (large-cap, in-domain) and the CSI 1000 (small-cap, out-of-domain) universes to assess generalization across asset pools. While both are temporally out-of-sample, the CSI 1000 evaluates zero-shot spatial generalization.

4.1.2 Baselines and Model Configuration. We compare Alpha-R1 against two classes of baselines. The first consists of traditional quantitative strategies, including statistical methods (PCA), tree-based ensembles (XGBoost and LightGBM), and deep reinforcement learning approaches (A2C and PPO). The second class comprises reasoning-enabled large language models, including Gemini 2.5 Pro Thinking, Claude 3.7 Sonnet Thinking, DeepSeek-R1, and the base model Qwen3-8B.

Table 1: Main Experiment Results. Performance comparison of Alpha-R1 against baselines across two asset pools (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 CSI 300 (In-Domain)				Asset Pool CSI 1000 (Out-of-Domain)			
		CR (%)	AR (%)	SR	MDD (%)	CR (%)	AR (%)	SR	MDD (%)
Non-LLM	Buy & Hold	3.03	6.70	0.33	10.49	9.64	22.14	0.80	16.87
	PCA	-0.48	0.40	-0.06	14.69	6.24	16.09	0.59	16.13
	XGBoost	-10.03	-21.65	-1.54	15.33	4.34	11.77	0.45	19.12
	LightGBM	-5.10	-10.26	-0.83	13.43	-5.37	-6.92	-0.26	23.88
	A2C	-5.52	-11.12	-0.85	11.22	11.80	26.30	1.15	14.00
	PPO	0.89	3.28	0.11	11.67	-6.44	-7.62	-0.25	29.31
LLM	Gemini 2.5 Pro Thinking	-7.04	-14.45	-1.01	15.08	-8.73	-15.38	-0.58	28.37
	Claude 3.7 Sonnet Thinking	-5.41	-10.23	-0.63	13.58	3.80	13.26	0.43	16.98
	DeepSeek-R1	-5.98	-11.93	-0.82	14.88	-7.58	-12.87	-0.50	27.89
	Qwen3-8B	-6.32	-12.41	-0.77	16.35	2.73	10.23	0.29	21.78
	Alpha-R1 (Ours)	12.99	27.59	1.62	6.76	42.49	78.18	4.03	9.25

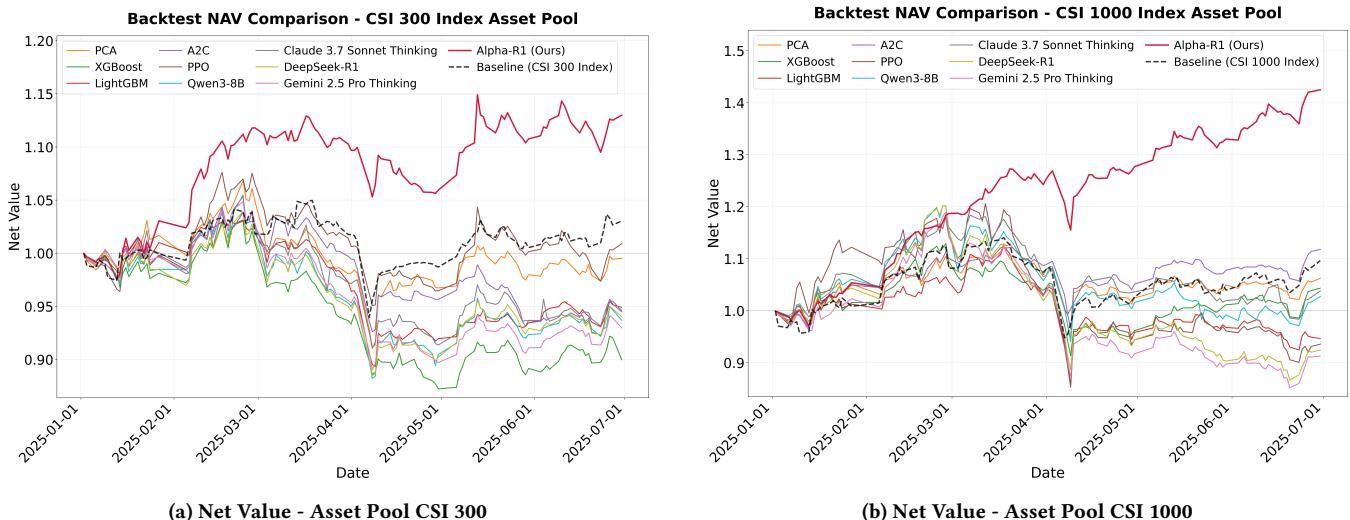


Figure 2: Cumulative Net Value Curves. Comparison of wealth accumulation trajectories on the 2025 testing set. (a) On the CSI 300, Alpha-R1 outperforms all baselines with lower drawdown. (b) On the Out-of-Domain CSI 1000, Alpha-R1 demonstrates superior zero-shot transferability, whereas traditional RL models exhibit inconsistent performance: A2C shows improved returns but with higher volatility, while PPO suffers from significant drawdowns, indicating overfitting and limited generalization.

Consistent with the execution protocol described in Section 3.5, we employ a *slot rotation strategy* with a holding period of $H = 5$ days, managing five concurrent sub-portfolios. Each slot selects $TopN = 10$ stocks, and trades are executed using 30-minute VWAP prices with a fixed bilateral transaction cost of 0.1%.

To prevent data leakage and ensure a fair evaluation, all benchmark LLMs have pre-training data cutoffs strictly no later than December 31, 2024. Alpha-R1 adopts Qwen3-8B as its backbone model. Inference is performed deterministically with `temperature=0` and `top_p=0.7`.

4.1.3 Evaluation Protocol. Performance is evaluated using cumulative return (CR), annualized return (AR), Sharpe ratio (SR), and maximum drawdown (MDD). 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 excess returns. Table 1 details the quantitative metrics, while Figure 2 visualizes the wealth accumulation trajectories.

4.2.1 Comparative Analysis on In-Domain Testing (CSI 300). As shown in Table 1 and Figure 2a, Alpha-R1 achieves the strongest performance on the CSI 300 domain.

- *Comparison with tree-based models.* Gradient boosting methods such as XGBoost, while effective on static prediction tasks, perform poorly in this setting (e.g., cumulative return of -10.03%). This highlights a critical limitation: in non-stationary markets, the dominant error source is often model misspecification rather than within-regime estimation. A purely dynamic model that re-estimates coefficients from limited and noisy samples induces high variance and can overreact to transient correlations as historical correlations deteriorate.
- *Comparison with reinforcement learning methods.* Deep RL approaches such as A2C and PPO struggle to adapt to the non-stationary market environment. A2C achieves a negative cumulative return of -5.52% with a Sharpe ratio of -0.85, while PPO shows only marginal positive returns (0.89%) with a low Sharpe ratio of 0.11. This suggests that numerical RL agents, despite their ability to learn sequential decision-making, are vulnerable to distributional shifts and fail to capture the semantic relationships that govern factor effectiveness across different market regimes.
- *Comparison with generic reasoning LLMs.* Models such as Claude 3.7 Sonnet and DeepSeek-R1 demonstrate strong general reasoning ability but lack domain-specific financial grounding. As a result, they fail to adequately account for risk constraints, leading to substantial drawdowns (approximately 15–16%) and negative Sharpe ratios.
- *Performance of Alpha-R1.* By interpreting our framework as a context-conditioned sparse linear model, Alpha-R1 attains a Sharpe ratio of 1.62 while limiting maximum drawdown to 6.76%. The fixed linear scorer provides a stable, low-variance mapping, while the LLM reasoning core functions as a gating network that activates or deactivates factors based on semantic alignment. This allows for stable and risk-aware decision-making even as market regimes shift.

4.2.2 Zero-Shot Generalization on Out-of-Domain (CSI 1000). The CSI 1000 experiment (Figure 2b) evaluates the model’s ability to generalize under a zero-shot setting. Compared with the CSI 300, the CSI 1000 represents a small-cap universe with higher volatility and more pronounced idiosyncratic dynamics.

Numerical reinforcement learning agents, such as A2C and PPO, exhibit limited generalization beyond the training domain. On the CSI 300, A2C achieves a negative cumulative return of -5.52% with a Sharpe ratio of -0.85, while PPO shows marginal positive returns (0.89%) with a low Sharpe ratio of 0.11. When transferred to the CSI 1000, A2C shows improved performance (cumulative return of 11.80%) but PPO suffers significant degradation with a negative return of -6.44% and a maximum drawdown of 29.31%, indicating sensitivity to domain-specific statistical patterns and overfitting issues.

In contrast, Alpha-R1 maintains strong performance on the out-of-domain CSI 1000 universe, achieving a cumulative return of 42.49% and a Sharpe ratio of 4.03. This confirms the advantage of delegating non-stationarity adaptation to the LLM. By conditioning

factor activation on richer state information (summarized from price dynamics and news narratives) and enforcing parsimonious selection, our approach reduces misspecification and estimation noise. This mechanism enables robust zero-shot transfer by capturing higher-level economic relationships rather than relying on noisy, domain-specific statistical regularities.

4.3 Ablation Study

To dissect the contributions of individual components, we conduct a series of ablation experiments to isolate the impact of key modules in Alpha-R1. All ablation variants are trained from scratch using the modified framework to ensure a fair comparison.

4.3.1 Ablation Variants. We consider the following four variants to evaluate the necessity of data inputs, factor representations, and the training objective:

- *w/o News:* Removes market news information from the state space S_t , forcing the model to rely solely on price data and factor values.
- *w/o Market Price:* Removes market price trends from S_t , retaining only news narratives and factor values.
- *w/o Semantic Description:* Replaces the natural language descriptions of factors with their raw mathematical formulas to test if the LLM relies on semantic reasoning or pattern matching.
- *w/o RL Optimization:* Uses the backbone model (Qwen3-8B) without the reinforcement learning alignment pipeline.

Table 2: Ablation Study Results. Performance contribution of different components (Testing Period: 2025.01.01 – 2025.06.30 on CSI 300).

Method	CR (%)	AR (%)	SR	MDD (%)
Alpha-R1 (Full)	12.99	27.59	1.62	6.76
Buy & Hold (CSI 300 Index)	3.03	6.70	0.33	10.49
w/o Market Price	10.24	22.42	1.24	12.87
w/o News	8.75	19.61	1.03	12.01
w/o Semantic Description	7.26	16.76	0.83	13.32
w/o RL Optimization	-6.32	-12.41	-0.77	16.35

4.3.2 Results and Analysis. Table 2 provides a quantitative decomposition of the performance contributions attributable to each module within the Alpha-R1 framework. By systematically ablating key components, we elucidate the specific mechanisms driving the model’s superior risk-adjusted returns. The empirical evidence supports three principal conclusions regarding the architecture’s efficacy:

The Imperative of Reinforcement Learning Alignment. The most profound performance disparity is observed between the full model and the w/o RL Optimization variant. The unaligned base model (Qwen3-8B) fails to generate positive alpha, yielding a Sharpe Ratio of -0.77, identical to the w/o RL Optimization case. This finding corroborates the hypothesis that general-purpose reasoning capabilities, while necessary, are insufficient for the stochastic and adversarial nature of financial markets. The RL alignment process

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

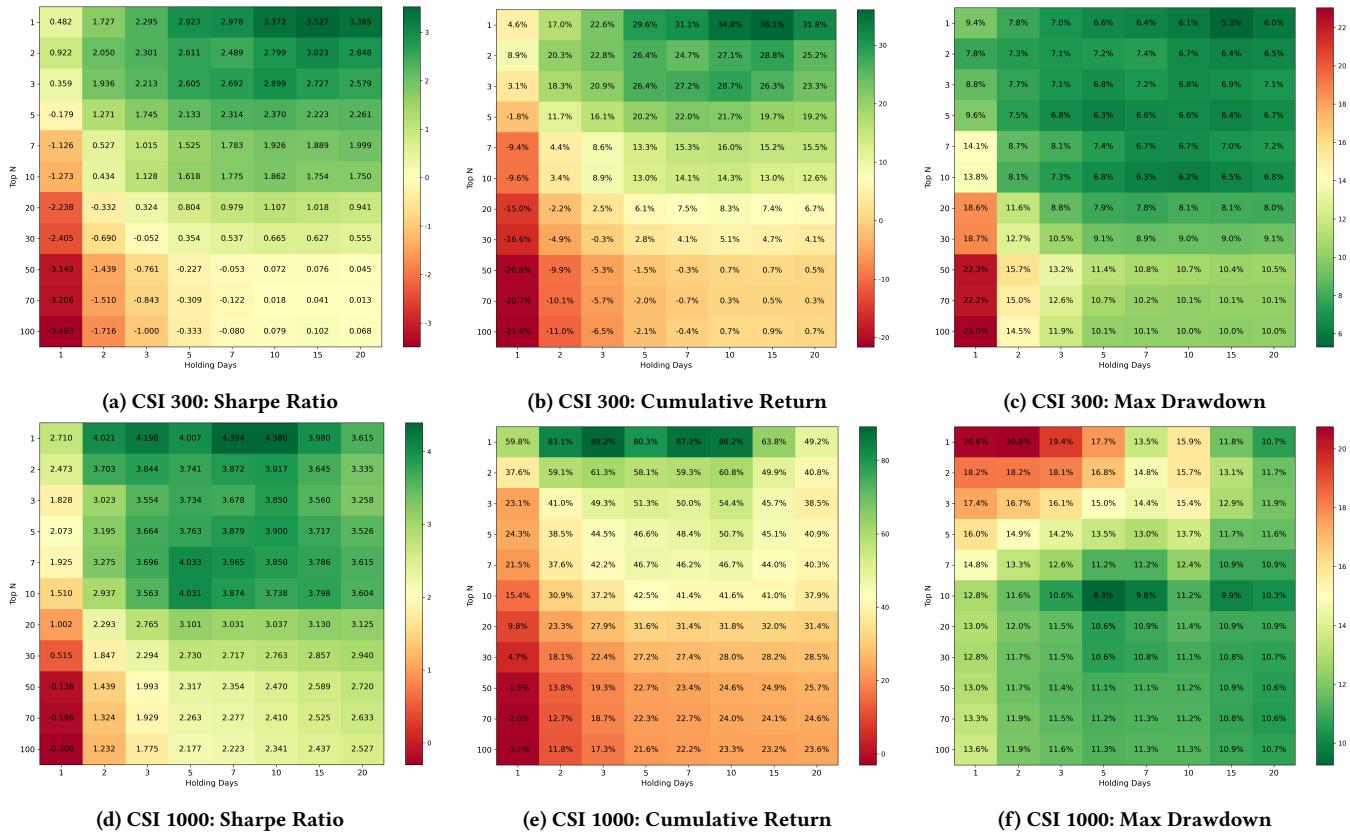


Figure 3: Parameter Sensitivity and Generalization Analysis. The heatmaps illustrate the impact of varying *TopN* and *Holding Days*. Green regions indicate favorable performance (High Sharpe/Return, Low Drawdown), while Red regions indicate poor performance. The broad Green clusters across both asset pools confirm the strategy’s robustness.

acts as a crucial transformation layer, grounding the LLM’s broad semantic knowledge into specific, risk-aware actionable policies. Without this targeted optimization, the model’s latent reasoning capacity cannot be effectively translated into profitable trading decisions.

Semantic Reasoning Over Mathematical Memorization. The degradation in performance observed in the w/o Semantic Description variant—where natural language factor descriptions are replaced by raw mathematical formulas—highlights the model’s reliance on semantic reasoning. The Sharpe Ratio declines by approximately 49% (from 1.62 to 0.83). This substantial gap suggests that Alpha-R1 does not merely memorize statistical patterns associated with factor formulas. Instead, it leverages the semantic richness of natural language to infer the economic rationale behind factors, enabling more robust selection in changing market regimes compared to purely symbolic processing.

Synergistic Integration of Multi-Modal Signals. The ablation of informational inputs reveals the complementary nature of news narratives and price dynamics. Excluding market news (w/o News) results in a more severe deterioration of the Sharpe Ratio (dropping to 1.03) compared to removing price trends (1.24). This implies that qualitative news data serves as a superior leading indicator

for detecting regime shifts than historical price momentum alone. However, the full model exhibits the lowest Maximum Drawdown (6.76%), significantly outperforming both data-truncated variants. This indicates a synergistic effect: while news provides high-level context, price dynamics offer granular trend confirmation, and their integration is essential for minimizing tail risk and ensuring robust execution.

In synthesis, the superior performance of Alpha-R1 is not attributable to any single isolated component but emerges from the holistic integration of semantic understanding, multi-modal context awareness, and rigorous RL-based alignment.

4.4 Semantic vs. Heuristic Gating Strategies

To validate the superiority of our semantic gating mechanism, we compare it against traditional heuristic gating strategies. We consider:

- **Lasso:** A classic sparse linear model selecting factors via L_1 regularization.
- **IC Momentum:** A momentum-based heuristic selecting the top 10 factors based on their recent average IC over a 20-day window.

Table 3: Gating Strategy Comparison. Performance of Alpha-R1 (Semantic Gating) versus heuristic gating methods (Lasso and IC Momentum) on the CSI 300 testing set.

Method	CR (%)	AR (%)	SR	MDD (%)
Alpha-R1 (Semantic Gating)	12.99	27.59	1.62	6.76
Lasso	1.58	4.63	0.20	11.12
IC Momentum	-6.33	-12.55	-0.80	13.29

As shown in Table 3, Alpha-R1 significantly outperforms both heuristic baselines. While Lasso achieves a small positive return (1.58%), it lags behind the semantic gating approach. IC Momentum performs poorly (-6.33% CR), likely because historical IC measures often fail to persist in highly dynamic regimes. In contrast, Alpha-R1’s semantic gating leverages market context to adaptively select factors, demonstrating superior robustness to regime shifts.

4.5 Robustness and Generalization

To assess the strategy’s resilience, we conduct a parameter sensitivity analysis visualized via heatmaps in Figure 3.

4.5.1 Parameter Sensitivity (CSI 300). The top row of Figure 3 visualizes performance metrics on the in-domain asset pool. We observe broad green regions indicating high Sharpe Ratios across a wide range of settings (e.g., Holding Days 3–10). Notably, performance remains stable even as *TopN* increases from 5 to 20. This suggests that Alpha-R1 identifies a cluster of effective factors rather than relying on a single outlier signal, thereby reducing concentration risk.

4.5.2 Zero-Shot Generalization (CSI 1000). The bottom row of Figure 3 highlights the model’s performance on the out-of-domain CSI 1000 asset pool. Despite the significant shift in asset characteristics, the heatmaps display a consistent distribution of high-performance green clusters similar to the in-domain testing set. This confirms that Alpha-R1 possesses strong zero-shot generalization capabilities. Unlike traditional models that degrade rapidly outside their training distribution, Alpha-R1 leverages its semantic understanding of market regimes to dynamically adjust factor selections for the new asset pool, delivering robust risk-adjusted returns without retraining.

5 Conclusion

This paper introduces Alpha-R1, a semantics-driven approach to quantitative investment that shifts the focus from static factor mining to context-aware reasoning. Rather than relying solely on historical correlations, Alpha-R1 employs a reinforcement-learning-trained large language model to reason over the economic rationale underlying factor performance and its dependence on evolving market conditions.

By constructing a dual-layer semantic context that integrates long-term market memory with real-time information, Alpha-R1 connects unstructured data sources with quantitative decision-making in a unified manner. Methodologically, we develop a market-aligned reinforcement learning framework based on Group Relative

Policy Optimization (GRPO), in which training is guided by objective market outcomes instead of subjective human feedback.

Extensive empirical results show that Alpha-R1 consistently outperforms traditional machine learning baselines and generic reasoning LLMs across multiple asset pools. Notably, Alpha-R1 demonstrates strong zero-shot generalization when transferred from the CSI 300 to the CSI 1000 universe, maintaining stable and profitable performance in a previously unseen, high-volatility environment, where conventional reinforcement learning agents experience marked degradation. These findings suggest that grounding factor selection in semantic reasoning and market feedback provides a viable approach for mitigating alpha decay and addressing non-stationarity in financial markets.

References

- [1] 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. arXiv:2503.21422 [q-fin.CP] <https://arxiv.org/abs/2503.21422>
- [2] Lang Cao. 2025. Chain-of-Alpha: Unleashing the Power of Large Language Models for Alpha Mining in Quantitative Trading. <https://arxiv.org/abs/2508.06312> arXiv:2508.06312
- [3] Mark M. Carhart. 1997. On Persistence in Mutual Fund Performance. *The Journal of Finance* 52, 1 (1997), 57–82. doi:10.1111/j.1540-6261.1997.tb03808.x
- [4] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémie Scheurer, Javier Rando, Rachel Freedman, Tomek Korbak, David Lindner, Pedro Freire, Tony Tong Wang, Samuel Marks, Charbel-Raphael Segerie, Mical Carroll, Andi Peng, Phillip J.K. Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththanjanan, Max Nadeau, Eric J Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Biyik, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. 2023. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *Transactions on Machine Learning Research* (2023). <https://arxiv.org/abs/2307.15217> Survey Certification, Featured Certification.
- [5] Chao Cheng, Bin Chen, Zhe Xiao, et al. 2024. Quantum Finance and Fuzzy Reinforcement Learning-Based Multi-Agent Trading System. *International Journal of Fuzzy Systems* 26 (2024), 2224–2245. doi:10.1007/s40815-024-01731-1
- [6] 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. arXiv:2508.13174 [cs.AI] <https://arxiv.org/abs/2508.13174>
- [7] Eugene F. Fama and Kenneth R. French. 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33, 1 (1993), 3–56. doi:10.1016/0304-405X(93)90023-5
- [8] Gang Feng, Stefano Giglio, and Dacheng Xiu. 2020. Taming the Factor Zoo: A Test of New Factors. *The Journal of Finance* 75, 3 (2020), 1327–1370. doi:10.1111/jofi.12883
- [9] Joachim Freyberger, Alejandro J. Salgado, and Andriy Shkliko. 2020. Dissecting Characteristics Nonparametrically. *The Review of Financial Studies* 33, 5 (2020), 2326–2377. <https://www.jstor.org/stable/48574462>
- [10] Shihao Gu, Bryan Kelly, and Dacheng Xiu. 2020. Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies* 33, 5 (2020), 2223–2273. doi:10.1093/rfs/hhaa009
- [11] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyi Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhushu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiaishi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qin, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qushu Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shangao Lu, Shangyan Zhou, Shanhua Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Lin, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin,

- X. Q. Li, Xiangyu Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature* 645, 8081 (Sept. 2025), 633–638. doi:10.1038/s41586-025-09422-z
- [12] 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*, Franck Dernoncourt, Daniel Preotiuc-Pietro, and Anastasia Shimorina (Eds.). Association for Computational Linguistics, Miami, Florida, US, 1028–1045. doi:10.18653/v1/2024.emnlp-industry.77
- [13] 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* (2025). https://arxiv.org/abs/2501.17992
- [14] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. *arXiv:2009.03300 [cs.CY]* https://arxiv.org/abs/2009.03300
- [15] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiaiwu 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*. https://arxiv.org/abs/2308.00352
- [16] 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*. https://arxiv.org/abs/2106.0985
- [17] 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*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 1827–1843. doi:10.18653/v1/2023.findings-emnlp.123
- [18] Zura Kakushadze. 2016. 101 formulaic alphas. *Wilmott* 2016, 84 (2016), 72–81. https://doi.org/10.1002/wilm.10525
- [19] Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2025. A Survey of Reinforcement Learning from Human Feedback. *Transactions on Machine Learning Research* (2025). https://arxiv.org/abs/2312.14925 Survey Certification.
- [20] 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)*. ACM, 4304–4315. doi:10.1145/3589334.3645611
- [21] 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*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 18517–18533. doi:10.18653/v1/2025.findings-emnlp.1005
- [22] 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*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 1755–1797. doi:10.18653/v1/2025.findings-naacl.96
- [23] 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*. https://arxiv.org/abs/2311.13743
- [24] Yuante Li, Xi 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*. https://arxiv.org/abs/2505.15155
- [25] 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. *arXiv:2309.03736 [q-fin.PM]* https://arxiv.org/abs/2309.03736
- [26] 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*. https://arxiv.org/abs/2307.10485
- [27] Zhaowei Liu, Xin Guo, Fangqi Lou, Lingfeng Zeng, Jinyi Niu, Zixuan Wang, Jiajie Xu, Weige Cai, Ziwei Yang, Xueqian Zhao, Chao Li, Sheng Xu, Dezhui Chen, Yun Chen, Zuo Bai, and Liwen Zhang. 2025. Fin-R1: A Large Language Model for Financial Reasoning through Reinforcement Learning. *arXiv:2503.16252 [cs.CL]* https://arxiv.org/abs/2503.16252
- [28] Alejandro Lopez-Lira and Yuehua Tang. 2023. Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. https://arxiv.org/abs/2304.07619 arXiv:2304.07619
- [29] Dakuai Lu, Hengkui Wu, Jiaqing Liang, Yipei Xu, Qianyu He, Yipeng Geng, Mengkun Han, Yingxiu Xin, and Yanghua Xiao. 2023. BBT-Fin: Comprehensive Construction of Chinese Financial Domain Pre-trained Language Model, Corpus and Benchmark. *arXiv:2302.09432 [cs.CL]* https://arxiv.org/abs/2302.09432
- [30] Jifang Mai, Shaohua Zhang, Haiqing Zhao, and Lijun Pan. 2024. Factor Investment or Feature Selection Analysis? *Mathematics* 13, 1 (2024), 9. doi:10.3390/math13010009
- [31] 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*. https://arxiv.org/abs/2405.14734
- [32] 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* (New Orleans, LA, USA) (NIPS '22). Curran Associates Inc., Red Hook, NY, USA, Article 2011, 15 pages. https://arxiv.org/abs/2203.02155
- [33] 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. *arXiv:2502.08127 [cs.CL]* https://arxiv.org/abs/2502.08127
- [34] Rafael Rafailev, 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*. https://arxiv.org/abs/2305.18290
- [35] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *arXiv:1707.06347 [cs.LG]* https://arxiv.org/abs/1707.06347
- [36] 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. DeepSeek-Math: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv:2402.03300 [cs.CL]* https://arxiv.org/abs/2402.03300
- [37] William F. Sharpe. 1964. Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. *The Journal of Finance* 19, 3 (1964), 425–442. doi:10.1111/j.1540-6261.1964.tb02865.x
- [38] Hao Shi, Weili Song, Xinting Zhang, Jiahé 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, Article 1392, 9 pages. doi:10.1609/aaai.v39i12.33365
- [39] 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). Association for Computing Machinery, New York, NY, USA, 2813–2822. doi:10.1145/3711896.3736838
- [40] Hariom Tatsat and Ariye Shater. 2025. Beyond the Black Box: Interpretability of LLMs in Finance. *arXiv:2505.24650 [cs.CE]* https://arxiv.org/abs/2505.24650
- [41] Robert Tibshirani. 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58, 1 (1996), 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x
- [42] 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*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 3120–3131. doi:10.18653/v1/2024.findings-acl.185
- [43] Saizhao Wang, Hang Yuan, Leon Zhou, Lionel Ni, Heung-Young 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*, Ivan Habernal, Peter Schulam, and Jörg Tiedemann (Eds.). Association for Computational Linguistics, Suzhou, China, 196–206. doi:10.18653/v1/2025.emnlp-demos.14
- [44] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabrowski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. BloombergGPT: A Large Language Model for Finance. *arXiv:2303.17564 [cs.LG]* https://arxiv.org/abs/2303.17564
- [45] Yijia Xiao, Edward Sun, Tong Chen, Fang Wu, Di Luo, and Wei Wang. 2025. Trading-R1: Financial Trading with LLM Reasoning via Reinforcement Learning.

- arXiv:2509.11420 [q-fin.TR] <https://arxiv.org/abs/2509.11420>
- [46] Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. 2025. TradingAgents: Multi-Agents LLM Financial Trading Framework. In *The First MARW: Multi-Agent AI in the Real World Workshop at AAAI 2025*. <https://arxiv.org/abs/2412.20138>
- [47] 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* (New Orleans, LA, USA) (*NIPS' 23*). Curran Associates Inc., Red Hook, NY, USA, Article 1454, 16 pages. <https://arxiv.org/abs/2306.05443>
- [48] Hang Yuan, Saizhuo Wang, and Jian Guo. 2024. Alpha-GPT 2.0: Human-in-the-Loop AI for Quantitative Investment. <https://arxiv.org/abs/2402.09746> arXiv:2402.09746
- [49] Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. 2023. Instruct-FinGPT: Financial Sentiment Analysis by Instruction Tuning of General-Purpose Large Language Models. arXiv:2306.12659 [cs.CL] <https://arxiv.org/abs/2306.12659>
- [50] 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* (Barcelona, Spain) (*KDD '24*). Association for Computing Machinery, New York, NY, USA, 4314–4325. doi:10.1145/3637528.3671801
- [51] Xuanyu Zhang, Qing Yang, and Dongliang Xu. 2023. XuanYuan 2.0: A Large Chinese Financial Chat Model with Hundreds of Billions Parameters. arXiv:2305.12002 [cs.CL] <https://arxiv.org/abs/2305.12002>