

# Navigating the Alpha Jungle: An LLM-Powered MCTS Framework for Formulaic Factor Mining

Yu Shi<sup>1</sup>, Yitong Duan<sup>1, 2</sup>, and Jian Li<sup>1</sup>

<sup>1</sup>Institute for Interdisciplinary Information Sciences, Tsinghua University

<sup>2</sup>Zhongguancun Institute of Artificial Intelligence

{shi-y23, dyt19}@mails.tsinghua.edu.cn, lapordge@gmail.com

## Abstract

Alpha factor mining is pivotal in quantitative investment for identifying predictive signals from complex financial data. While traditional formulaic alpha mining relies on human expertise, contemporary automated methods, such as those based on genetic programming or reinforcement learning, often struggle with search inefficiency or yield alpha factors that are difficult to interpret. This paper introduces a novel framework that integrates Large Language Models (LLMs) with Monte Carlo Tree Search (MCTS) to overcome these limitations. Our framework leverages the LLM's instruction-following and reasoning capability to iteratively generate and refine symbolic alpha formulas within an MCTS-driven exploration. A key innovation is the guidance of MCTS exploration by rich, quantitative feedback from financial backtesting of each candidate factor, enabling efficient navigation of the vast search space. Furthermore, a frequent subtree avoidance mechanism is introduced to enhance search diversity and prevent formulaic homogenization, further improving performance. Experimental results on real-world stock market data demonstrate that our LLM-based framework outperforms existing methods by mining alphas with superior predictive accuracy and trading performance. The resulting formulas are also more amenable to human interpretation, establishing a more effective and efficient paradigm for formulaic alpha mining.

## 1 Introduction

Predicting price movements in financial markets, characterized by low signal-to-noise ratios, remains a central challenge in quantitative investment. A common strategy to enhance model predictiveness is the extraction of predictive signals, or alpha factors (hereafter simply “alphas”), from stock data (Qian, Hua, and Sorensen 2007; Tulchinsky 2019). Current alpha factor mining methodologies broadly fall into two categories: neural network-based and formula-based. Neural approaches (e.g., FactorVAE (Duan et al. 2022), HIST (Xu et al. 2021a), REST (Xu et al. 2021b)) implicitly construct complex alphas via deep learning, capturing intricate patterns but often suffering from a lack of interpretability. In contrast, formula-based methods aim to discover alphas represented by explicit mathematical expressions. These alpha factors are traditionally human-crafted, reflecting market insights (e.g., Fama-French factors (Fama and French 1992), financial anomalies (Harvey, Liu, and Zhu 2016; Hou, Xue, and Zhang 2020)). In recent years, automated techniques have emerged,

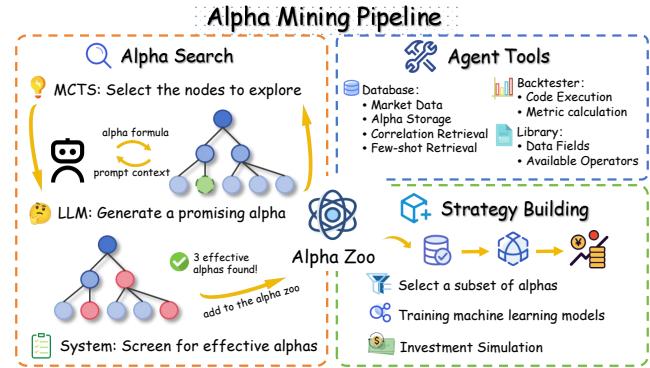


Figure 1: A high-level schematic of our proposed alpha mining pipeline. The pipeline features an iterative Alpha Search loop where an LLM, guided by MCTS, generates and refines formulas. Effective alphas are collected in an Alpha Zoo before being used in the final Strategy Building stage.

employing methods like genetic programming or reinforcement learning to discover such formulaic alphas (Shi et al. 2024; Yu et al. 2023; Zhang et al. 2020, 2023b).

Despite their promise, existing automated formulaic alpha mining approaches face significant limitations. **First, the discovered alphas often exhibit poor interpretability.** These automated methods frequently engage in unconstrained, data-driven exploration of the vast alpha space, often without sufficient guidance from financial theory or domain expertise. Consequently, the resulting formulas can be overly complex and opaque. This lack of transparency poses considerable challenges in practical investment scenarios: it hinders practitioners' ability to understand the underlying economic rationale of a strategy, makes it difficult to attribute portfolio performance accurately, and can erode trust, thereby impeding the adoption of these alphas even if they show promise in backtests. **Second, current methodologies often suffer from search inefficiency.** The search for a sufficient number of effective alpha factors typically requires generating and evaluating an enormous volume of candidate formulas. This exhaustive search process, while necessary due to the low signal density, inherently increases the likelihood of discovering spurious relationships and overfitting to the training data (Harvey, Liu, and Zhu 2016). As a result, many discovered alphas may exhibit poor generalization and deliver underwhelming out-of-sample performance.

Addressing the identified shortcomings necessitates innovative approaches. In this light, Large Language Models (LLMs) emerge as a promising direction, given their vast prior knowledge and strong reasoning capabilities which are well-suited for generating interpretable alphas—a potential demonstrated in analogous tasks like financial investment (Yu et al. 2024) and code generation (Li et al. 2024a). Drawing inspiration from advancements in LLM reasoning (e.g., Chain-of-Thought (Wei et al. 2022), Tree-of-Thought (Yao et al. 2024)) and the efficacy of Monte Carlo Tree Search (MCTS) (Coulom 2007; Silver et al. 2016) in enhancing LLM performance on complex problems (Zhang et al. 2024), we frame alpha mining as an MCTS-driven search problem. Within this framework, each node in the tree represents a candidate alpha formula, allowing for a systematic exploration and iterative refinement within the vast and complex alpha space. Figure 1 provides a high-level illustration of this entire pipeline.

Unlike tasks such as mathematical derivation, where evaluating the contribution of intermediate steps towards the final solution is often challenging before the derivation is complete, alpha factor mining provides fine-grained feedback on each candidate alpha formula through backtesting. We leverage this detailed feedback to guide our search. We initiate the search with a LLM-generated alpha formula, as the root node of the search tree. Then we utilize the LLM to iteratively refine and improve the formulas, expanding the tree with new, potentially superior nodes. Furthermore, to mitigate the homogeneity of generated alpha formulas, we conduct frequent subtree mining on effective alphas and explicitly instruct the LLM to avoid using the most frequent subtrees during generation. By explicitly diversifying the search away from these common motifs, our approach enhances search efficiency and improves the quality of discovered alphas.

The synergy between MCTS and LLMs has indeed shown promise in various reasoning tasks (Zhao, Lee, and Hsu 2023; DeLorenzo et al. 2024). Our work is distinct in its specific application of this synergy to the unique challenges of formulaic alpha mining. Unlike general reasoning tasks that use MCTS to explore a set of predefined actions or have the LLM evaluate abstract states (Xie et al. 2024; Li et al. 2025; Dainese et al. 2024), our framework leverages the LLM as a *generative prior* for symbolic alpha formulas. Crucially, the MCTS exploration is guided by rich, quantitative, and domain-specific feedback from financial backtesting performed on each candidate alpha. This iterative loop—where the LLM’s generative capabilities are steered by MCTS informed by empirical financial performance—offers a distinct advantage.

The main contributions of this paper can be summarized as follows:

- We propose an LLM-Powered MCTS framework for formulaic alpha mining, modeling the task as a tree search-based reasoning problem where the LLM performs multi-step formula refinement guided by detailed backtesting feedback.
- We design a frequent subtree avoidance method to improve search efficiency and alpha effectiveness by guiding

the LLM to explore less common yet potentially effective formula structures.

- We conduct a series of experiments to demonstrate the effectiveness of our proposed framework. The alphas mined by our framework achieve superior prediction performance while maintaining good interpretability, compared to those from other methods.

## 2 Preliminary

### Alpha Factor Mining

We consider a financial market with  $n$  stocks observed over  $T$  trading days. For each stock  $i \in \{1, \dots, n\}$  and day  $t \in \{1, \dots, T\}$ , its state is described by a feature vector  $\mathbf{x}_{i,t} \in \mathbb{R}^m$ . Raw features include daily open, high, low, close prices (OHLC), trading volume, and Volume-Weighted Average Price (VWAP). The complete market history is a tensor  $\mathbf{X} \in \mathbb{R}^{T \times n \times m}$ . Correspondingly, future returns are organized in a matrix  $\mathbf{Y} \in \mathbb{R}^{T \times n}$ , where  $y_{i,t}$  is the realized future return for stock  $i$  subsequent to day  $t$ . To capture temporal patterns, we use a lookback window of length  $\tau$ . An *alpha factor*,  $f$ , maps the historical feature data for this window,  $\mathbf{X}_{t-\tau+1:t} = \{\mathbf{X}_s \mid t - \tau < s \leq t\}$ , to a vector of predictive scores  $\mathbf{v}_t = f(\mathbf{X}_{t-\tau+1:t}) \in \mathbb{R}^n$ . Each  $v_{i,t}$  represents the alpha’s assessment of stock  $i$ ’s future return.

Alpha factor mining aims to discover a diverse set of  $K$  alphas,  $\mathcal{F} = \{f_1, \dots, f_K\}$ , by searching within the vast space of all possible alpha factors, denoted as  $\mathcal{A}$ . The outputs of these individual alphas,  $\{\mathbf{v}_{k,t} = f_k(\mathbf{X}_{t-\tau+1:t})\}_{k=1}^K$ , are typically aggregated by a combination model,  $g$ , into a composite alpha vector  $\mathbf{z}_t = g(\{\mathbf{v}_{k,t}\}_{k=1}^K; \theta_g)$ , where  $\theta_g$  are model parameters. The quality of this composite signal is evaluated using a predefined performance metric,  $\mathcal{L}$  (e.g., Information Coefficient). Optimal parameters for the combination model,  $\theta_g^*$ , are learned by maximizing this metric:

$$\theta_g^*(\mathcal{F}) = \arg \max_{\theta_g} \mathcal{L}(g(\{\mathbf{v}_{k,t}\}_{k=1}^K; \theta_g), \mathbf{Y}). \quad (1)$$

Let  $g^*(\mathcal{F})$  be the combination model with parameters  $\theta_g^*(\mathcal{F})$ . The overarching goal is to find an optimal set of alpha factors  $\mathcal{F}^* \subset \mathcal{A}$  that maximizes the performance of this optimally combined signal:  $\mathcal{F}^* = \arg \max_{\mathcal{F} \subset \mathcal{A}} \mathcal{L}(g^*(\mathcal{F}), \mathbf{Y})$ . This constitutes a challenging bilevel optimization problem due to the immense search space  $\mathcal{A}$  and the complex interactions between alpha factors.

### Formulaic Alpha

In this work, we focus on *formulaic alphas*: alpha factors defined by mathematical expressions. These expressions are constructed from operators and operands. Operands typically include raw input features (e.g.,  $\text{close}_{i,t}$ ) and numerical constants. Operators apply mathematical transformations; for example, time-series operators can be used to construct an alpha like  $\text{Ma}(\text{close}, 5) - \text{Ma}(\text{close}, 20)$ . This specific alpha captures a price trend by contrasting short-term with long-term moving averages of closing prices. A complete list of available operators is provided in Appendix E. Formulaic alphas are naturally represented as expression trees (leaf nodes: operands; internal nodes: operators), making their structured

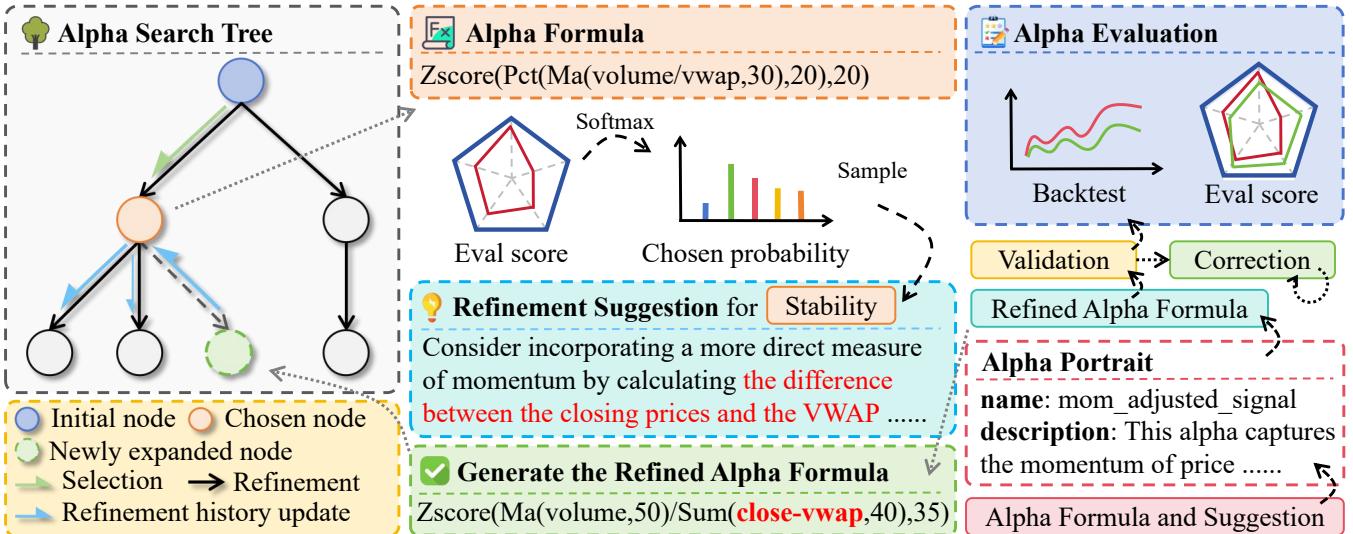


Figure 2: Overview of our LLM-powered MCTS framework. The process begins with node selection via UCT. A refinement dimension is then chosen based on the node’s multi-dimensional evaluation scores. The LLM first proposes a conceptual refinement suggestion for that dimension and then translates it into a concrete formula. The new formula is backtested, and its performance results are used to expand the tree with a new node.

yet flexible nature amenable to the automated mining techniques central to our framework.

### 3 Methodology

Our proposed alpha mining framework integrates LLMs with MCTS to automate the discovery and refinement of alpha factors. The framework’s objective is to search within the vast space of possible alpha formulas, denoted as  $\mathcal{A}$ , to find a set of high-performing alphas. Figure 2 provides a conceptual illustration. The core iterative process involves: (1) selecting a promising node (alpha formula) using the Upper Confidence Bound for Trees (UCT) criterion (Kocsis and Szepesvári 2006); (2) expanding this node by having the LLM generate a refined alpha, guided by performance feedback on specific evaluation dimensions; and (3) evaluating the new alpha via backtesting, with results forming a new node in the search tree  $\mathcal{T}$  (see Appendix C for an illustrative case). The LLM’s role is twofold: first, to propose targeted refinement suggestions, and second, to translate these suggestions into a concrete alpha formula  $f \in \mathcal{A}$ . Iteratively, high-performing alphas that satisfy a set of predefined criteria  $\mathcal{C}$  (e.g.,  $IC > 0.02$ ) are collected into an effective alpha repository,  $\mathcal{F}_{zoo}$ .

#### Selection

The selection step in our MCTS framework navigates the exploration-exploitation trade-off. Each node  $s \in \mathcal{T}$  represents an alpha, characterized by its formula  $f_s$  and refinement history. An action  $a$  corresponds to a specific refinement applied to  $s$ . Each state-action pair  $(s, a)$  maintains a quality value  $Q(s, a)$ , representing the maximum reward (alpha score) observed in the subtree rooted at the child node resulting from this action. We employ the UCT criterion to select

the optimal action  $a^*$ :

$$a^* = \arg \max_{a \in A(s)} \left( Q(s, a) + c \sqrt{\frac{\ln(N_s)}{N_{s'}}} \right) \quad (2)$$

where  $A(s)$  is the set of existing actions from state  $s$ ,  $N_s$  is the visit count of the parent state  $s$ ,  $N_{s'}$  is the visit count of the child state  $s' = \text{child}(s, a)$ , and  $c$  is the exploration weight.

Unlike standard MCTS, which expands only leaf nodes, our approach allows any node to be selected for expansion. This is crucial for iteratively refining promising, but not terminal, alpha ideas. To enable this, we augment the action space  $A(s)$  at any internal node  $s$  with a “virtual” expansion action,  $a_e$ . The selection process thus considers the full action set  $A(s) \cup \{a_e\}$ . The UCT score for this virtual action is computed by adapting Equation 2, where we define a virtual visit count for the prospective new node  $s'_e$  as  $N_{s'_e} = 1 + |C(s)|$ , with  $C(s)$  being the set of existing children of  $s$ . If  $a_e$  is selected, node  $s$  is chosen for expansion. This mechanism ensures that promising, non-leaf nodes can be further refined.

#### Expansion

Upon selecting a node  $s$  for expansion, a new, refined alpha factor  $f_{new}$  is generated. This process is structured to enhance the LLM’s effectiveness and the quality of refinements.

**Dimension-Targeted Refinement Suggestion.** Each node  $s$  is associated with a multi-dimensional evaluation score vector  $\mathbf{E}_s = [e_1, \dots, e_q] \in [0, e_{\max}]^q$ . To guide refinement towards areas of weakness while maintaining explorative diversity, we stochastically select a target dimension  $i^*$  for improvement. The probability of choosing dimension  $i$  is defined as:

$$P(i^* = i | s) = \text{Softmax} ((e_{\max} \cdot \mathbf{1}_q - \mathbf{E}_s) / T)_i \quad (3)$$

where  $\mathbf{1}_q$  is a  $q$ -dimensional vector of ones and  $T$  is a temperature parameter. This strategy prioritizes dimensions with lower scores. Once a dimension  $i^*$  is selected, the LLM generates a textual refinement suggestion  $d_{s,i^*}$  aimed at improving performance on that dimension. This is framed as a few-shot learning task, where the context includes effective alphas from  $\mathcal{F}_{zoo}$  (see Appendix D).

**Alpha Formula Generation and Validation.** Following the targeted suggestion, we employ a two-step generation process. First, the LLM articulates the refined conceptual hypothesis, which is then used to prompt the generation of the concrete formula. This process, ensuring the formula aligns with a clear investment rationale, can be formalized as:

$$d_{s,i^*} \sim p_{LLM}(\cdot | s, i^*, \mathcal{F}_{zoo}) \quad (4)$$

$$f_{new} \sim p_{LLM}(\cdot | d_{s,i^*}, f_s) \quad (5)$$

The generated formula  $f_{new}$  undergoes an automated validation check,  $\text{IsValid}(f_{new})$ . If invalid, feedback is provided to the LLM for iterative correction. A valid formula and its evaluation results constitute the new node  $s_{new}$  in the MCTS tree.

### Multi-Dimensional Alpha Evaluation

The evaluation of a candidate alpha  $f$  is performed directly via backtesting, bypassing the simulation phase of traditional MCTS. A key challenge is the evolving nature of the effective alpha repository  $\mathcal{F}_{zoo}$ , which progressively raises the bar for new alphas. To address this, we employ a relative ranking approach. The rank of  $f$  against the repository for a given metric  $m$  is:

$$R(f, m, \mathcal{F}_{zoo}) = \frac{1}{|\mathcal{F}_{zoo}|} \sum_{f' \in \mathcal{F}_{zoo}} \mathbb{I}(m(f) < m(f')) \quad (6)$$

where  $\mathbb{I}(\cdot)$  is the indicator function. This provides an adaptive evaluation criterion, avoiding fixed thresholds that may be too stringent early on or too lenient later.

To provide granular feedback for refinement, we evaluate  $f$  across a set of  $q$  dimensions  $\mathcal{D}$ , exemplified here by **Effectiveness, Stability, Turnover, Diversity, and Overfitting Risk** (detailed in Appendix D). For each dimension  $D_i \in \mathcal{D} \setminus \{\text{Overfitting Risk}\}$ , we compute a score  $e_i$  based on its percentile rank using a corresponding metric  $m_i$ :

$$e_i(f) = 1 - R(f, m_i, \mathcal{F}_{zoo}) \quad (7)$$

The assessment of **Overfitting Risk**, denoted  $e_{\text{overfit}}$ , is distinct. We leverage an LLM that analyzes the formula  $f$  and its refinement history  $H(s)$ , providing a qualitative judgment:  $e_{\text{overfit}} = \text{LLM}_{\text{eval}}(f, H(s))$  (see Appendix K for the specific prompt). The overall alpha score, which serves as the reward signal for MCTS, is the aggregate of these dimensional scores:

$$S(f) = \frac{1}{|\mathcal{D}|} \sum_{i=1}^q e_i(f) \quad (8)$$

This score  $S(f)$  is used to update the Q-values in the search tree.

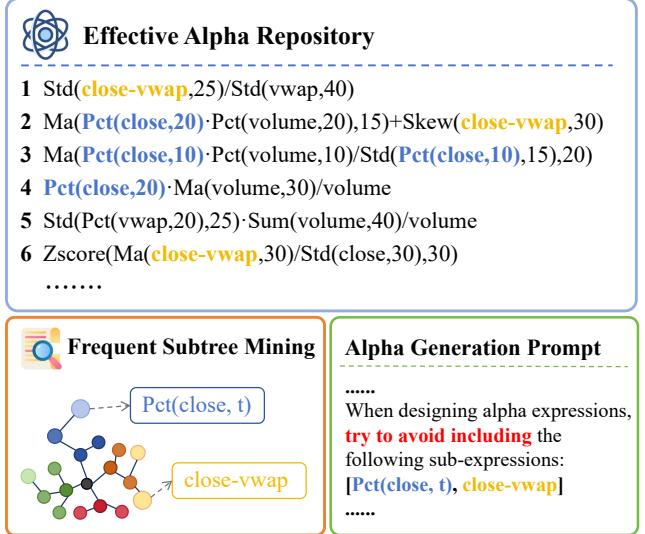


Figure 3: Illustration of the Frequent Subtree Avoidance (FSA). The set of effective alphas from the Alpha Repository is mined for frequent subtrees. The most frequent ones are identified, and the LLM is subsequently instructed to avoid generating new formulas containing these common structural motifs.

### Backpropagation

During backpropagation, the reward  $S(f_{new})$  of the newly evaluated alpha at node  $s_{new}$  updates the statistics of all nodes along the path from the root to its parent. For each ancestor node  $s_k$  on this path, leading to its child  $s_{k+1}$  via action  $a_k$ , we perform the following updates:

$$N_{s_k} \leftarrow N_{s_k} + 1 \quad (9)$$

$$Q(s_k, a_k) \leftarrow \max(Q(s_k, a_k), S(f_{new})) \quad (10)$$

This ensures that the value of a path reflects the best outcome discovered within its entire subtree. Crucially, to enhance the quality of subsequent refinement suggestions and the accuracy of overfitting assessment, the LLM is provided with rich contextual information: the refinement history of the current node's parent, children, and siblings. This allows the LLM to analyze the refinement trajectory and avoid redundant suggestions.

### Frequent Subtree Avoidance

To mitigate alpha formula homogenization and prevent the over-exploitation of common motifs, we introduce Frequent Subtree Avoidance (FSA), inspired by the concept of “root genes” in AutoAlpha (Zhang et al. 2020). We define a *root gene* as a subtree in an alpha’s expression tree whose leaves are exclusively raw input features (e.g., ‘close’, ‘high’). To focus on structure, we use an operator  $\text{Abs}(\cdot)$  that abstracts away concrete parameter values from an expression tree. For instance,  $\text{Abs}(\text{Ma}(\text{vwap}, 20))$  becomes  $\text{Ma}(\text{vwap}, t)$ . The set of abstracted root genes for an alpha  $f$  is denoted  $\mathcal{G}(f)$ .

The FSA mechanism operates by first identifying frequent closed root genes from the repository  $\mathcal{F}_{zoo}$ . A root gene is “closed” if none of its immediate supertrees share the same support count, which helps identify maximal common

patterns. The support for an abstracted root gene  $\bar{g}$  is:

$$\text{Support}(\bar{g}) = \frac{1}{|\mathcal{F}_{zoo}|} \sum_{f' \in \mathcal{F}_{zoo}} \mathbb{I}(\bar{g} \subseteq \bar{\mathcal{G}}(f')) \quad (11)$$

We select the top- $k$  most frequent closed root genes to form a set of forbidden structures,  $\mathcal{G}_{\text{forbidden}}$ . During generation (Eq. 5), the LLM is constrained to produce a new formula  $f_{\text{new}}$  that avoids these motifs:

$$\text{Constraint: } \bar{\mathcal{G}}(f_{\text{new}}) \cap \mathcal{G}_{\text{forbidden}} = \emptyset \quad (12)$$

FSA acts as a regularization on the generation process. By discouraging common structures, it guides the MCTS search towards more structurally diverse and potentially novel regions of the alpha space  $\mathcal{A}$ , enabling a more efficient exploration, as demonstrated in Section 4.

## 4 Experiment

We evaluate our proposed framework on real-world stock market data, addressing the following research questions (RQs):

- Q1:** How does our approach compare to baselines in predictive performance?
- Q2:** Are MCTS and Frequent Subtree Avoidance effective components within our framework?
- Q3:** How does the interpretability of alpha formulas mined by our method compare to others?

### Experiment Settings

**Data** Our experiments are conducted on the Chinese A-shares market. To ensure comprehensive market representation, our experiments separately target two stock pools: the CSI 300 Index (large-cap, liquid stocks) and the CSI 1000 Index (small to mid-cap stocks). We define two distinct prediction targets: the 10-day return and the 30-day return of the stocks, with buying and selling executed at the closing price. The dataset is split chronologically into training (2011/01/01–2020/12/31) and testing (2021/01/01–2024/11/30) periods.

**Baselines for Comparison** We compare our framework with several formulaic alpha mining methods. **DSO** (Deep Symbolic Optimization) (Landajuela et al. 2022) is a deep learning framework for symbolic optimization. **GP** employs genetic programming for alpha mining. **AlphaGen** (Yu et al. 2023) is a reinforcement learning framework for formulaic alpha mining. **AlphaForge** (Shi et al. 2024) features a generative-predictive architecture; to ensure a fair comparison of generative capabilities, we use only its alpha mining network. Among LLM-based approaches, **CoT** (Chain-of-Thought) (Wei et al. 2022) prompts LLMs for step-by-step reasoning to directly generate alpha factors. **ToT** (Tree-of-Thought) (Yao et al. 2024) enables LLMs to explore diverse reasoning paths in a tree structure. Lastly, **FAMA** (Li et al. 2024b) leverages LLMs with in-context examples to diversify formulas and a “chain-of-experience” to learn from past successes.

We use OpenAI’s GPT4.1 model as the LLM for both our method and the LLM-based baselines. Furthermore, we evaluate the performance of our framework when using different

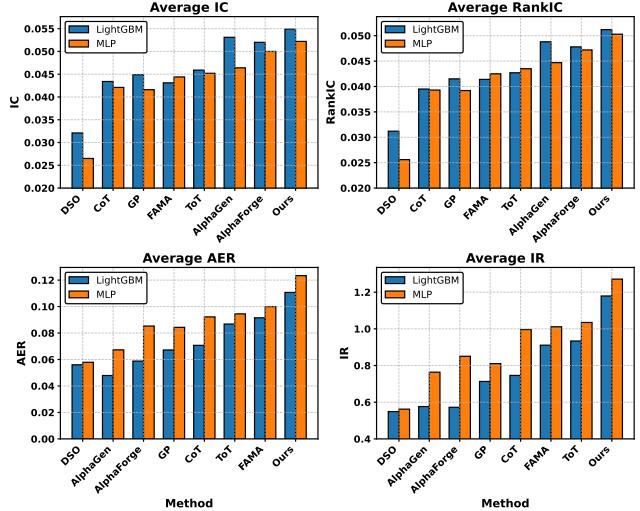


Figure 4: The average predictive performance of LightGBM and MLP models trained on alphas mined by different methods.

LLM backbones, as detailed in Appendix H. Regarding the potential data leakage issue in LLMs, please refer to the discussion in Appendix H.

To ensure a fair and rigorous comparison of algorithmic efficiency, we benchmark all methods based on a controlled “search count” (i.e., the number of unique alpha formulas generated and evaluated). This metric normalizes for the vast differences in computational cost per generation step across methods (e.g., a single LLM call vs. a GP mutation) and provides a direct measure of search space exploration efficiency. This approach is well-suited because our framework and all baselines inherently involve a distinct search process, where each iteration yields a new candidate alpha formula. For LLM-based methods (including ours), we report the best performance achieved with search counts of 1,000, 2,000, or 3,000. For other methods, the search count is incrementally increased from a small value until performance converges, capped at 600,000 (200× the LLM-based methods’ maximum). This experimental design facilitates two key comparisons: first, it allows for an equitable assessment of our framework against other LLM-based methods under similar, well-defined computational budgets; second, it enables a robust comparison of search efficiency against other non-LLM-based methods.

### Experiment 1: Prediction Performance Comparison

We evaluate the predictive performance of alphas generated by our method against baselines. To comprehensively assess the effectiveness of the generated alpha sets, we employ two representative machine learning models: LightGBM (Ke et al. 2017), a popular gradient boosting framework known for its efficiency, and a 3-layer Multi-Layer Perceptron (MLP), which can capture complex non-linear relationships. For each alpha generation method, we create alpha sets of three distinct sizes—10, 50, and 100—to serve as input features for these models. This allows for a thorough comparison of the mined alpha sets across varying sizes. Both input alphas

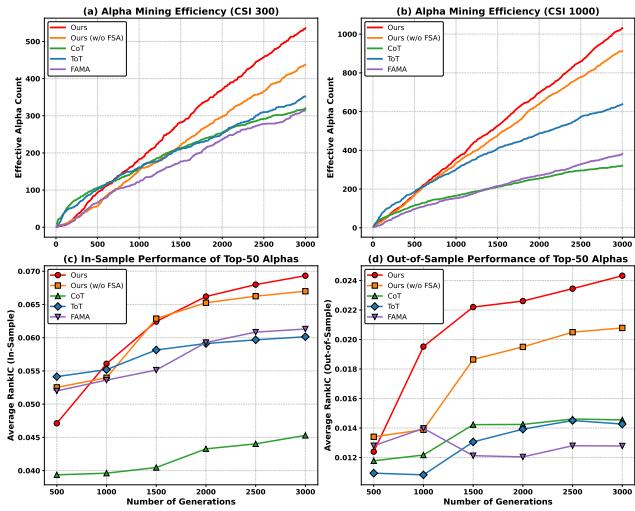


Figure 5: Analysis of search dynamics. (a, b) Alpha mining efficiency, measured as the count of effective alphas found versus total generated. (c, d) Average in-sample and out-of-sample RankIC of the top 50 alphas over generations.

and target returns undergo cross-sectional rank normalization before training to mitigate outlier influence. The predictive power of the alphas is evaluated using two standard metrics in quantitative finance: the Information Coefficient (IC) and the Rank Information Coefficient (RankIC).

To assess the practical profitability of the mined alphas in simulated real-world stock market scenarios, we follow established evaluation methodologies (Yu et al. 2023) and conduct backtests. Specifically, we employ a top- $k$ /drop- $n$  portfolio construction strategy, implemented on the Qlib platform (Yang et al. 2020). The detailed configuration of this strategy is described in Appendix G. The practical trading performance is evaluated using two key metrics: Annualized Excess Return (AER), which measures the strategy’s profitability, and Information Ratio (IR), which quantifies its risk-adjusted performance.

The combination of these four metrics (IC, RankIC, AER, and IR) provides a comprehensive evaluation of the mined alphas. As illustrated in Figure 4 (with detailed results presented in Appendix I), our framework consistently outperforms baselines across all metrics. This demonstrates that the alphas mined by our framework possess superior predictive capabilities for future stock returns, which can be effectively translated into trading profitability. These findings are further corroborated by our experiments on the U.S. stock market, with detailed results presented in Appendix H.

## Experiment 2: Ablation Study

We conduct ablation studies to evaluate three key components of our framework: MCTS, multi-dimensional feedback, and FSA.

Table 1 presents the impact of these components on predictive performance. When incorporating Effectiveness and Diversity as feedback, MCTS demonstrates superior predictive performance over CoT and ToT. Performance progressively improves with the integration of additional feedback dimensions. Notably, while Turnover feedback slightly reduces IC

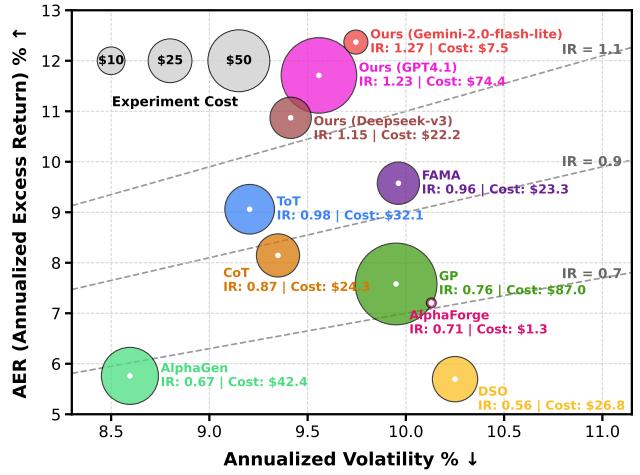


Figure 6: Cost-performance analysis across various methods. The plot shows Annualized Excess Return (AER) vs. Annualized Volatility (lower is better). Bubble size indicates the estimated single-run cost. Dashed lines represent constant Information Ratio (IR) levels.

and RankIC, it enhances practical trading metrics (AER, IR) by mitigating transaction costs. The integration of FSA yields further improvements across all metrics for both LightGBM and MLP models. These results underscore the individual and collective contributions of MCTS, multi-dimensional feedback, and FSA to our framework’s efficacy.

Beyond predictive performance, we analyze the search dynamics in Figure 5. Subplots (a) and (b) assess search efficiency by plotting the number of effective alphas mined against the total generated alphas. Our framework, even without FSA, demonstrates higher search efficiency than other LLM-based methods. FSA further amplifies this advantage. Furthermore, subplots (c) and (d) evaluate the quality and generalization of the top 50 alphas (selected by in-sample RankIC) over generations. The in-sample performance of our method’s alphas improves as the search progresses (c). More importantly, this improvement translates to superior out-of-sample performance (d), indicating that our framework discovers more generalizable alphas compared to baselines.

To establish a more standardized and equitable basis for comparison, we present a cost-performance analysis in Figure 6. We estimate the cost of a single experimental run for each method by unifying its runtime and API usage into a monetary value based on public cloud computing prices (detailed in Appendix H). The analysis reveals that our framework achieves a favorable risk-return profile, with its variants occupying the highest Information Ratio (IR) contours. Notably, the overall cost and performance of our framework are primarily driven by the choice of the underlying LLM. For instance, employing a lightweight model like Gemini-2.0-flash-lite yields a high IR of 1.27 at a minimal cost of \$7.5. In contrast, using the more powerful GPT-4.1 results in a slightly lower IR of 1.23 at a substantially higher cost of \$74.4. This highlights that our framework offers the flexibility to select an appropriate LLM, enabling a desirable balance between performance and computational budget.

Table 1: Ablation study of our framework’s components. Best results are highlighted in **bold**. We ablate across five evaluation dimensions: Effectiveness (Eff.), Diversity (Div.), Turnover (Turn.), Stability (Stab.), and Overfitting Risk (O.R.). The checkmark (✓) indicates a dimension is included in the search, while the cross (✗) indicates it is not.

Search Strategy	Included Evaluation Dimensions					LightGBM				MLP			
	Eff.	Div.	Turn.	Stab.	O.R.	IC	RankIC	AER	IR	IC	RankIC	AER	IR
CoT	✓	✓	✗	✗	✗	0.0434	0.0395	0.0707	0.7461	0.0421	0.0393	0.0922	0.9962
ToT	✓	✓	✗	✗	✗	0.0459	0.0427	0.0868	0.9337	0.0452	0.0435	0.0945	1.0348
MCTS	✓	✗	✗	✗	✗	0.0409	0.0374	0.0941	0.9775	0.0400	0.0376	0.0935	1.0010
MCTS	✓	✓	✗	✗	✗	0.0501	0.0476	0.1003	1.0106	0.0486	0.0462	0.1023	1.0462
MCTS	✓	✓	✓	✗	✗	0.0492	0.0457	0.1063	1.1062	0.0489	0.0462	0.1185	1.2556
MCTS	✓	✓	✓	✓	✗	0.0495	0.0462	0.1030	1.0331	0.0491	0.0465	0.1093	1.1773
MCTS	✓	✓	✓	✓	✓	0.0515	0.0479	0.1075	1.1121	0.0503	0.0478	0.1166	1.2127
MCTS+FSA	✓	✓	✓	✓	✓	<b>0.0549</b>	<b>0.0512</b>	<b>0.1107</b>	<b>1.1792</b>	<b>0.0522</b>	<b>0.0503</b>	<b>0.1234</b>	<b>1.2712</b>

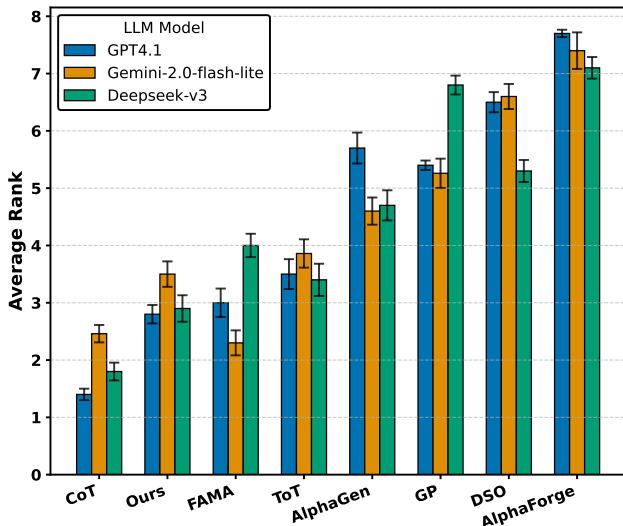


Figure 7: Interpretability Ranking Comparison, showing mean and standard deviation over 5 random seeds.

### Experiment 3: Interpretability of Alpha Formulas

In this experiment, we evaluate the interpretability of alpha formulas mined by different methods. We define the interpretability of an alpha formula by its capacity to articulate a reasonable logic, a specific market phenomenon, or an investment strategy. To quantify this, we randomly select one alpha formula per method and employ LLMs to rank their interpretability. We repeat this process 50 times and compute the average rank for each method. To mitigate potential biases from a single LLM, we aggregate rankings from three distinct LLMs. The results are presented in Figure 7. The findings indicate that formulas mined by our framework exhibit interpretability second only to those generated by the CoT method. This result is particularly noteworthy when compared to non-LLM baselines, whose formulas were consistently ranked as less interpretable. This suggests that our approach achieves a compelling trade-off, delivering strong predictive performance while maintaining a high degree of interpretability.

Furthermore, acknowledging that LLMs might inherently favor formulas generated through LLM-driven processes,

we provide illustrative examples of alpha formulas from our method and non-LLM-based approaches in Appendix H. While a formal human study is outside the scope of this work, a qualitative inspection of these examples reveals a clear difference: formulas from our method tend to have more discernible logic compared to the often opaque structures produced by non-LLM methods.

## 5 Related Work

Automated formulaic alpha mining has traditionally relied on genetic programming (GP) frameworks (Lin et al. 2019; Zhang et al. 2020), with other notable approaches employing reinforcement learning (Yu et al. 2023) and deep learning-based generative models (Shi et al. 2024). More recently, LLMs have been introduced, with approaches like FAMA (Li et al. 2024b) and AlphaAgent (Tang et al. 2025) leveraging them for direct alpha generation guided by in-context examples or heuristics. In contrast, our method frames alpha discovery as a formal reasoning task, uniquely employing MCTS to systematically explore the structured space of mathematical formulas. This approach is inspired by recent advancements in tree search-based reasoning, such as Tree of Thoughts (ToT) (Yao et al. 2024) and RethinkMCTS (Li et al. 2024a), which enhance LLM planning by structuring generation as a search process. A further literature review is provided in Appendix A.

## 6 Conclusion

We introduce an LLM-Powered Monte Carlo Tree Search (MCTS) framework for formulaic alpha mining. This approach models alpha mining as a tree search, where an LLM iteratively generates and refines candidate formulas, critically guided by quantitative feedback from financial backtesting. To foster search efficiency and alpha effectiveness, we incorporate a Frequent Subtree Avoidance mechanism. Experimental results demonstrate that our framework mines alphas with superior predictive accuracy and trading performance, while also offering enhanced interpretability and search efficiency compared to existing methods. This work pioneers a promising direction for leveraging LLMs and MCTS to tackle the complex challenge of automated formulaic alpha mining in finance.

## References

- Amihud, Y.; and Mendelson, H. 1986. Asset pricing and the bid-ask spread. *Journal of financial Economics*, 17(2): 223–249.
- Coulom, R. 2007. Monte-Carlo Tree Search in Crazy Stone,”. In *Proc. Game Prog. Workshop, Tokyo, Japan*, 74–75.
- Cui, C.; Wang, W.; Zhang, M.; Chen, G.; Luo, Z.; and Ooi, B. C. 2021. Alphaevolve: A learning framework to discover novel alphas in quantitative investment. In *Proceedings of the 2021 International conference on management of data*, 2208–2216.
- Dainese, N.; Merler, M.; Alakuijala, M.; and Marttinen, P. 2024. Generating code world models with large language models guided by monte carlo tree search. *arXiv preprint arXiv:2405.15383*.
- De Bondt, W. F.; and Thaler, R. 1985. Does the stock market overreact? *The Journal of finance*, 40(3): 793–805.
- DeLorenzo, M.; Chowdhury, A. B.; Gohil, V.; Thakur, S.; Karri, R.; Garg, S.; and Rajendran, J. 2024. Make every move count: Llm-based high-quality rtl code generation using mcts. *arXiv preprint arXiv:2402.03289*.
- Duan, Y.; Wang, L.; Zhang, Q.; and Li, J. 2022. Factorvae: A probabilistic dynamic factor model based on variational autoencoder for predicting cross-sectional stock returns. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 4468–4476.
- Efficiency, S. M. 1993. Returns to buying winners and selling losers: Implications for. *The Journal of Finance*, 48(1): 65–91.
- Fama, E. F.; and French, K. R. 1992. The cross-section of expected stock returns. *the Journal of Finance*, 47(2): 427–465.
- French, K. R.; Schwert, G. W.; and Stambaugh, R. F. 1987. Expected stock returns and volatility. *Journal of financial Economics*, 19(1): 3–29.
- Harvey, C. R.; Liu, Y.; and Zhu, H. 2016. ... and the cross-section of expected returns. *The Review of Financial Studies*, 29(1): 5–68.
- Hou, K.; Xue, C.; and Zhang, L. 2020. Replicating anomalies. *The Review of financial studies*, 33(5): 2019–2133.
- Hu, Z.; Liu, C.; Feng, X.; Zhao, Y.; Ng, S.-K.; Luu, A. T.; He, J.; Koh, P. W.; and Hooi, B. 2024. Uncertainty of Thoughts: Uncertainty-Aware Planning Enhances Information Seeking in Large Language Models. *arXiv preprint arXiv:2402.03271*.
- Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; and Liu, T.-Y. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.
- Kocsis, L.; and Szepesvári, C. 2006. Bandit based monte-carlo planning. In *European conference on machine learning*, 282–293. Springer.
- Kou, Z.; Yu, H.; Luo, J.; Peng, J.; and Chen, L. 2024. Automate strategy finding with llm in quant investment. *arXiv preprint arXiv:2409.06289*.
- Landajuela, M.; Lee, C. S.; Yang, J.; Glatt, R.; Santiago, C. P.; Aravena, I.; Mundhenk, T.; Mulcahy, G.; and Petersen, B. K. 2022. A unified framework for deep symbolic regression. *Advances in Neural Information Processing Systems*, 35: 33985–33998.
- Li, Q.; Xia, W.; Du, K.; Dai, X.; Tang, R.; Wang, Y.; Yu, Y.; and Zhang, W. 2024a. RethinkMCTS: Refining Erroneous Thoughts in Monte Carlo Tree Search for Code Generation. *arXiv preprint arXiv:2409.09584*.
- Li, S.; Dong, S.; Luan, K.; Di, X.; and Ding, C. 2025. Enhancing Reasoning through Process Supervision with Monte Carlo Tree Search. *arXiv preprint arXiv:2501.01478*.
- Li, Z.; Song, R.; Sun, C.; Xu, W.; Yu, Z.; and Wen, J.-R. 2024b. Can Large Language Models Mine Interpretable Financial Factors More Effectively? A Neural-Symbolic Factor Mining Agent Model. In *Findings of the Association for Computational Linguistics ACL 2024*, 3891–3902.
- Lin, X.; Chen, Y.; Li, Z.; and He, K. 2019. Stock Alpha Mining Based On Genetic Algorithm. Technical report, Huatai Securities Research Center.
- Qian, E. E.; Hua, R. H.; and Sorensen, E. H. 2007. *Quantitative equity portfolio management: modern techniques and applications*. Chapman and Hall/CRC.
- Ren, T.; Zhou, R.; Jiang, J.; Liang, J.; Wang, Q.; and Peng, Y. 2024. RiskMiner: Discovering Formulaic Alphas via Risk Seeking Monte Carlo Tree Search. In *Proceedings of the 5th ACM International Conference on AI in Finance*, 752–760.
- Shi, H.; Song, W.; Zhang, X.; Shi, J.; Luo, C.; Ao, X.; Arian, H.; and Seco, L. 2024. AlphaForge: A Framework to Mine and Dynamically Combine Formulaic Alpha Factors. *arXiv preprint arXiv:2406.18394*.
- Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. 2016. Mastering the game of Go with deep neural networks and tree search. *nature*, 529(7587): 484–489.
- Tang, Z.; Chen, Z.; Yang, J.; Mai, J.; Zheng, Y.; Wang, K.; Chen, J.; and Lin, L. 2025. AlphaAgent: LLM-Driven Alpha Mining with Regularized Exploration to Counteract Alpha Decay. *arXiv preprint arXiv:2502.16789*.
- Tulchinsky, I. 2019. *Finding Alphas: A quantitative approach to building trading strategies*. John Wiley & Sons.
- Wang, C.; Deng, Y.; Lyu, Z.; Zeng, L.; He, J.; Yan, S.; and An, B. 2024a. Q\*: Improving multi-step reasoning for llms with deliberative planning. *arXiv preprint arXiv:2406.14283*.
- Wang, P.; Li, L.; Shao, Z.; Xu, R.; Dai, D.; Li, Y.; Chen, D.; Wu, Y.; and Sui, Z. 2024b. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9426–9439.
- Wang, S.; Yuan, H.; Ni, L. M.; and Guo, J. 2024c. Quantagent: Seeking holy grail in trading by self-improving large language model. *arXiv preprint arXiv:2402.03755*.

- Wang, S.; Yuan, H.; Zhou, L.; Ni, L. M.; Shum, H.-Y.; and Guo, J. 2023. Alpha-gpt: Human-ai interactive alpha mining for quantitative investment. *arXiv preprint arXiv:2308.00016*.
- Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.
- Xie, Y.; Goyal, A.; Zheng, W.; Kan, M.-Y.; Lillicrap, T. P.; Kawaguchi, K.; and Shieh, M. 2024. Monte carlo tree search boosts reasoning via iterative preference learning. *arXiv preprint arXiv:2405.00451*.
- Xu, W.; Liu, W.; Wang, L.; Xia, Y.; Bian, J.; Yin, J.; and Liu, T.-Y. 2021a. Hist: A graph-based framework for stock trend forecasting via mining concept-oriented shared information. *arXiv preprint arXiv:2110.13716*.
- Xu, W.; Liu, W.; Xu, C.; Bian, J.; Yin, J.; and Liu, T.-Y. 2021b. Rest: Relational event-driven stock trend forecasting. In *Proceedings of the web conference 2021*, 1–10.
- Yang, X.; Liu, W.; Zhou, D.; Bian, J.; and Liu, T.-Y. 2020. Qlib: An ai-oriented quantitative investment platform. *arXiv preprint arXiv:2009.11189*.
- Yang, X.; Yang, X.; Fang, S.; Xian, B.; Li, Y.; Wang, J.; Xu, M.; Pan, H.; Hong, X.; Liu, W.; et al. 2025. R&d-agent: Automating data-driven ai solution building through llm-powered automated research, development, and evolution. *arXiv preprint arXiv:2505.14738*.
- Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.; Cao, Y.; and Narasimhan, K. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.
- Yu, S.; Xue, H.; Ao, X.; Pan, F.; He, J.; Tu, D.; and He, Q. 2023. Generating synergistic formulaic alpha collections via reinforcement learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 5476–5486.
- Yu, Y.; Li, H.; Chen, Z.; Jiang, Y.; Li, Y.; Zhang, D.; Liu, R.; Suchow, J. W.; and Khashanah, K. 2024. FinMem: A performance-enhanced LLM trading agent with layered memory and character design. In *Proceedings of the AAAI Symposium Series*, volume 3, 595–597.
- Zhang, D.; Huang, X.; Zhou, D.; Li, Y.; and Ouyang, W. 2024. Accessing gpt-4 level mathematical olympiad solutions via monte carlo tree self-refine with llama-3 8b. *arXiv preprint arXiv:2406.07394*.
- Zhang, S.; Chen, Z.; Shen, Y.; Ding, M.; Tenenbaum, J. B.; and Gan, C. 2023a. Planning with large language models for code generation. *arXiv preprint arXiv:2303.05510*.
- Zhang, T.; Li, Y.; Jin, Y.; and Li, J. 2020. Autoalpha: an efficient hierarchical evolutionary algorithm for mining alpha factors in quantitative investment. *arXiv preprint arXiv:2002.08245*.
- Zhang, T.; Zhang, Z. A.; Fan, Z.; Luo, H.; Liu, F.; Liu, Q.; Cao, W.; and Jian, L. 2023b. OpenFE: automated feature generation with expert-level performance. In *International Conference on Machine Learning*, 41880–41901. PMLR.
- Zhao, Z.; Lee, W. S.; and Hsu, D. 2023. Large language models as commonsense knowledge for large-scale task planning. *Advances in Neural Information Processing Systems*, 36: 31967–31987.
- Zhou, A.; Yan, K.; Shlapentokh-Rothman, M.; Wang, H.; and Wang, Y.-X. 2023. Language agent tree search unifies reasoning acting and planning in language models. *arXiv preprint arXiv:2310.04406*.

## A Summary of Appendix

This appendix provides extensive supplementary material to support the methodologies, experiments, and findings presented in the main paper. The contents are organized as follows:

- **Section B: Further Related Work.** A review of related work in automated formulaic alpha mining and tree search-based reasoning.
- **Section C: Illustrative Case Study.** A step-by-step walkthrough of our framework’s MCTS workflow.
- **Section D: Method Details.** Elaboration on the generation and refinement of alpha formulas, the multi-dimensional evaluation metrics, and the dynamic search budget strategy.
- **Section E: List of Operators.** A comprehensive list of operators employed within the alpha formulas.
- **Section F: Pseudo-code.** The algorithmic pseudo-code of our framework.
- **Section G: Experimental Setup Details.** Details on the dataset, hyperparameter configurations, model specifications, backtesting strategy, and performance metrics.
- **Section H: Extended Experimental Results and Analyses.** Includes comparative analyses against other baselines, comparison results on the U.S. stock market, investigations into data leakage and LLM sensitivity, and qualitative analyses of alpha characteristics and interpretability.
- **Section I: Full Experiment Results.** Comprehensive results of the prediction performance comparison experiments.
- **Section J: Limitations.** A discussion of the limitations of our framework.
- **Section K: LLM Agent Prompts.** Details of key prompts used for alpha generation, refinement, and overfitting risk assessment.

## B Further Related Work

### Automated Formulaic Alpha Mining

**Traditional Methods** As mentioned in the main text, traditional alpha mining encompasses several paradigms. Genetic programming (GP) is a foundational approach. Early work like GPLearn (Lin et al. 2019) applies GP with a pre-defined set of time-series operators, while later enhancements like AutoAlpha (Zhang et al. 2020) improve efficiency by initializing the search with effective, low-depth alphas. AlphaEvolve (Cui et al. 2021) further advances this line of work by evolving computation graphs instead of simple formula trees. Beyond GP, other paradigms include AlphaGen (Yu et al. 2023), which uses reinforcement learning to optimize an entire alpha set, and AlphaForge (Shi et al. 2024), which designs a deep learning-based generative-predictive structure.

**Other LLM-based Frameworks** Several other works have explored the application of LLMs in quantitative finance, though they fall outside the scope of our direct comparison. For instance, AlphaGPT (Wang et al. 2023) focuses on a human-AI interaction paradigm, using prompt engineering to translate researcher insights into alphas. A framework by Kou et al. (Kou et al. 2024) mines alphas from multimodal financial data via a multi-agent system. Similarly, QuantAgent (Wang et al. 2024c) introduces a two-loop system where an LLM refines strategies from a knowledge base updated through real-world testing. These methods are excluded from our comparative benchmarks due to their reliance on non-standard data (e.g., multimodal), interactive human feedback, or methodologies that are not sufficiently detailed or open-sourced for a fair and reproducible comparison.

### Tree Search-based Reasoning

As mentioned in the main text, tree search methods are increasingly used to maximize the exploration capabilities of LLMs, allowing for different levels of search and planning (Zhang et al. 2023a; Hu et al. 2024). They are now widely applied in LLM-based agents and reasoning tasks (Wang et al. 2024b,a). LATS (Zhou et al. 2023), for example, extends this concept by viewing the LLM as a general agent that conducts exploration at both the reasoning and action levels, which is conceptually related to our approach of building a formula through a sequence of operator selections.

## C Illustrative Case Study of MCTS Workflow

To elucidate the workflow of our proposed MCTS framework for alpha factor mining, this section provides a detailed step-by-step example.

Initially, as depicted in Figure 8, an initial alpha formula is generated by the LLM. After undergoing a multi-dimensional evaluation, this formula becomes the root node ( $V_0$ ) of the MCTS search tree. For this illustrative case, the target search count (initial search budget) is initialized to 3. This target is dynamically incremented by 1 (budget increment) each time a newly generated node achieves a new high score, encouraging deeper exploration of promising search trees.

Following the construction of the root node, the MCTS process proceeds with node expansion. As shown in Figure 9, the root node  $V_0$  is expanded to generate its first child node,  $V_1$ . The dimension for refinement is selected by sampling according to Equation (3) (in this instance, the *Stability* dimension is chosen). Subsequently, the LLM generates targeted refinement

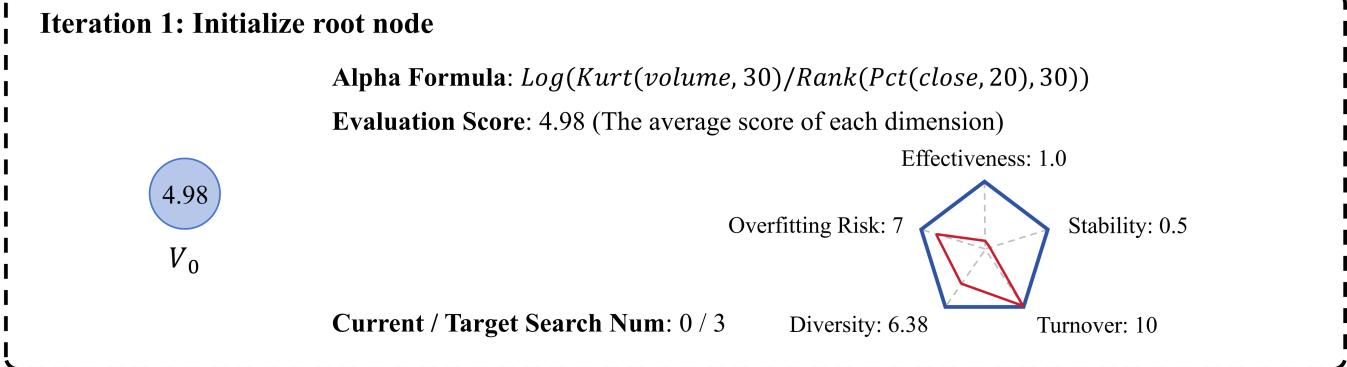


Figure 8: Case Study: MCTS root node generation.

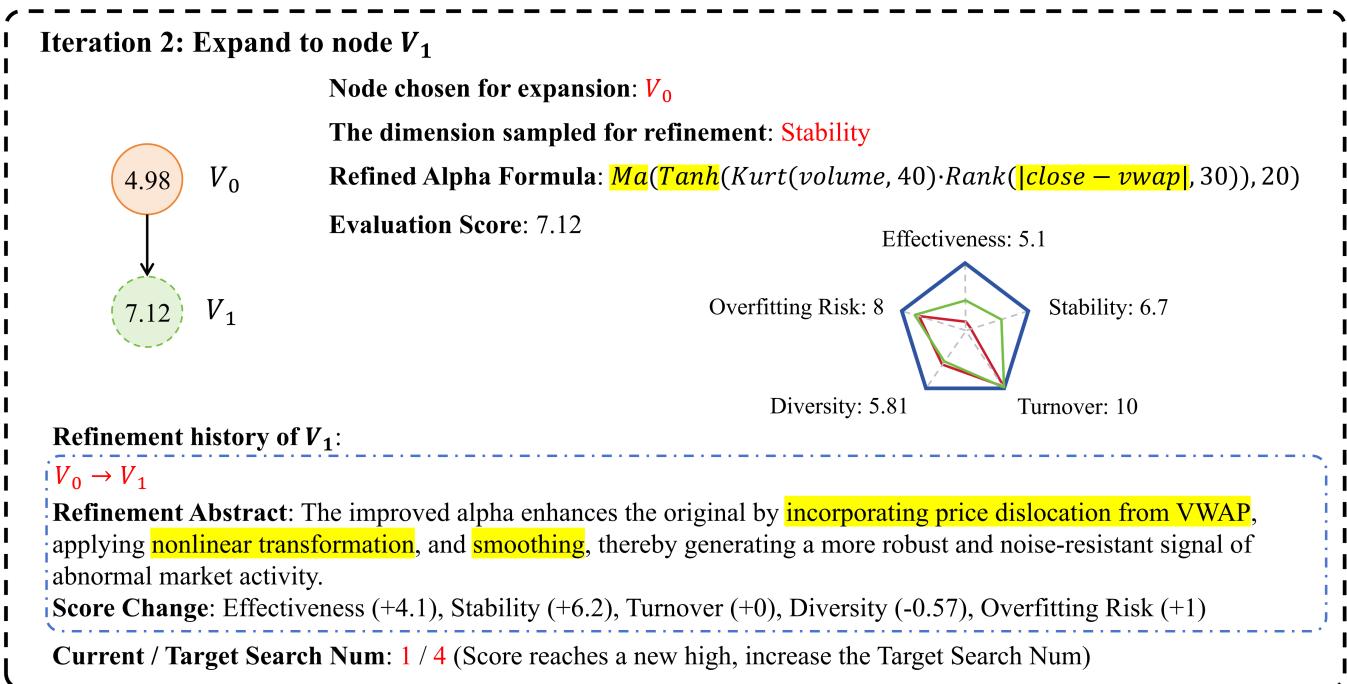


Figure 9: Case Study: expansion to node  $V_1$ .

suggestions. Based on these suggestions, a refined alpha formula is produced, with the modifications highlighted within the formula representation. This refined alpha formula is then subjected to the same multi-dimensional evaluation process, yielding the scored node  $V_1$ . The refinement history for  $V_1$  is updated to include a summary of this specific refinement step and the corresponding change in evaluation scores. Finally, both the target search count (if  $V_1$ 's score is a new high) and the current search count are updated.

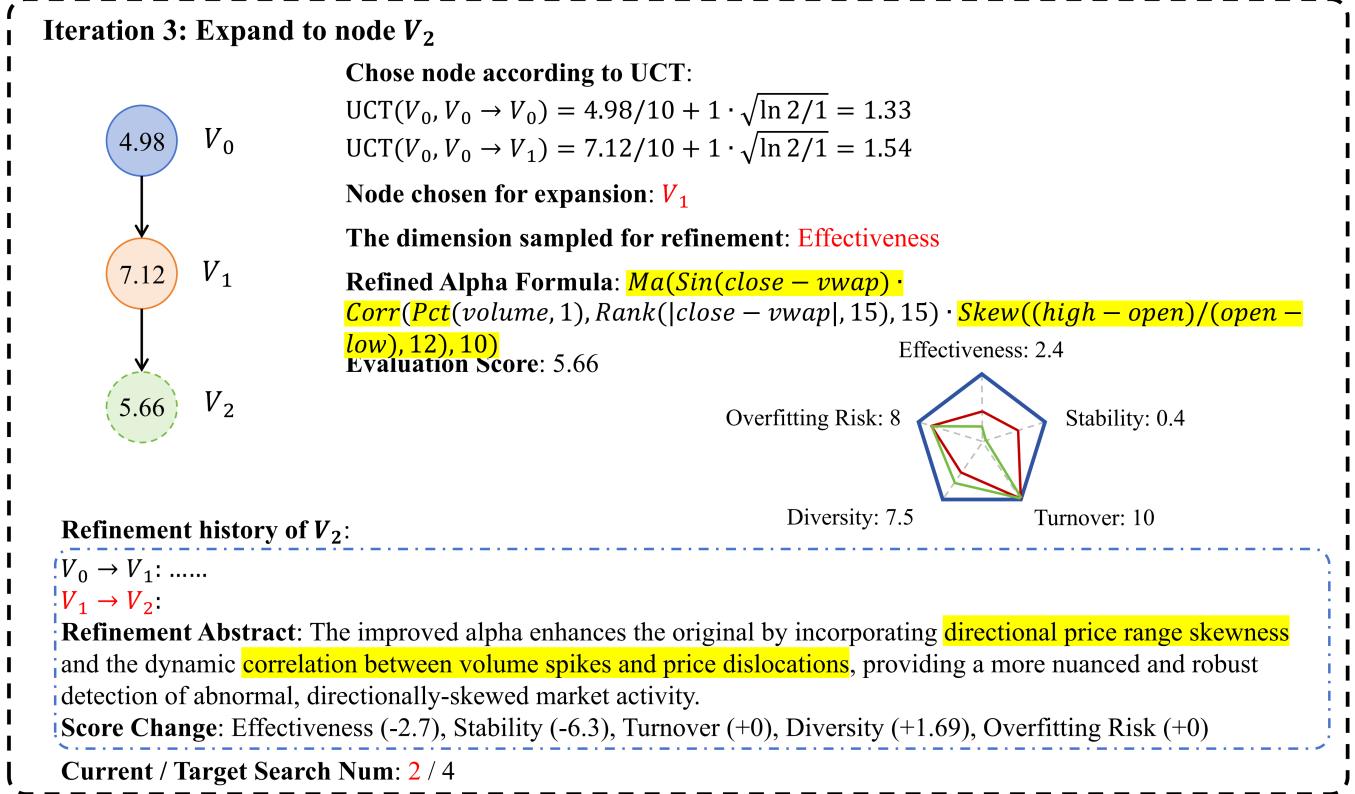


Figure 10: Case Study: expansion to node  $V_2$ .

The expansion process continues, as illustrated in Figure 10. Based on the UCT criterion, node  $V_1$  is selected for the next expansion. Similar to the previous step, a dimension for refinement is sampled. The LLM then generates an improved alpha formula based on this dimension, which is subsequently evaluated to create node  $V_2$ . Notably, the refinement history of  $V_2$  is cumulative, encapsulating all refinement steps and score evolutions from the root node  $V_0$  through  $V_1$  to  $V_2$ . The current search count is then incremented.

Figures 11 and 12 showcase the outcomes of the two subsequent expansion steps, leading to the generation of nodes  $V_3$  and  $V_4$ , respectively. Upon the expansion that creates node  $V_4$ , the current search count reaches the predetermined target. At this point, the expansion phase for the current MCTS iteration concludes. All five generated nodes ( $V_0, \dots, V_4$ ) are then systematically reviewed. Nodes that meet predefined quality and performance criteria are added to the effective alpha repository.

This sequence completes one full iteration of the MCTS-driven alpha mining process. Our framework is designed to iteratively repeat this entire cycle, enabling the efficient and systematic mining of effective alphas.

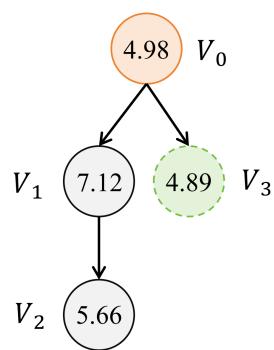
## D Method Details

In this section, we elaborate on specific details of our framework, focusing on the generation process, evaluation metrics, and the search budget allocation strategy.

### Alpha Formula Generation

We decompose the generation of alpha formulas into a two-step process: first, an *alpha portrait* is generated, and subsequently, the corresponding alpha formula is derived from this portrait. An alpha portrait is a textual representation that includes the alpha's name, a concise description of its underlying investment logic, and its formula expressed in pseudo-code. This two-step approach decouples the conceptual design of an alpha from its concrete construction, thereby reducing the instruction-following complexity for the LLM and enhancing the quality of the generated formulas.

#### Iteration 4: Expand to node $V_3$



**Chose node according to UCT:**

$$UCT(V_0, V_0 \rightarrow V_0) = 4.98/10 + 1 \cdot \sqrt{\ln 2/1} = 1.33$$

$$UCT(V_0, V_0 \rightarrow V_1) = \max(7.12, 5.66)/10 + 1 \cdot \sqrt{\ln 2/2} = 1.30$$

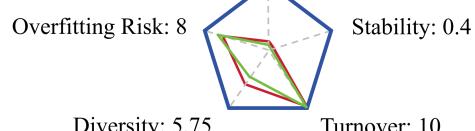
**Node chosen for expansion:**  $V_0$

**The dimension sampled for refinement:** Stability

**Refined Alpha Formula:**  $\text{Sign}(\text{Tanh}(\text{Ma}((\text{high} - \text{open})/(\text{high} + \text{open}), 10))) \cdot \text{Zscore}(\text{volume}, 14) \cdot \text{Rank}(|\text{close} - \text{vwap}|, 50))$

**Evaluation Score:** 4.89

Effectiveness: 0.3



**Refinement history of  $V_3$ :**

$V_0 \rightarrow V_1$ : .....

$V_0 \rightarrow V_3$ :

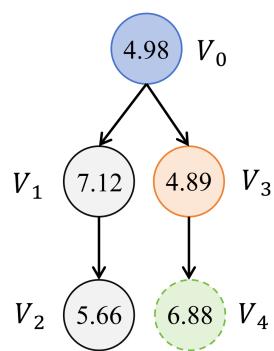
**Refinement Abstract:** The improved alpha factor enhances the original by incorporating intraday volatility and price dispersion to better isolate abnormal volume spikes specifically during turbulent and dislocated market conditions.

**Score Change:** Effectiveness (-0.7), Stability (-0.1), Turnover (+0), Diversity (-0.63), Overfitting Risk (+1)

**Current / Target Search Num:** 3 / 4

Figure 11: Case Study: expansion to node  $V_3$ .

#### Iteration 5: Expand to node $V_4$



**Chose node according to UCT:**

$$UCT(V_0, V_0 \rightarrow V_0) = 4.98/10 + 1 \cdot \sqrt{\ln 2/2} = 1.09$$

$$UCT(V_0, V_0 \rightarrow V_1) = \max(7.12, 5.66)/10 + 1 \cdot \sqrt{\ln 2/2} = 1.30$$

$$UCT(V_0, V_0 \rightarrow V_3) = 4.89/10 + 1 \cdot \sqrt{\ln 2/1} = 1.32$$

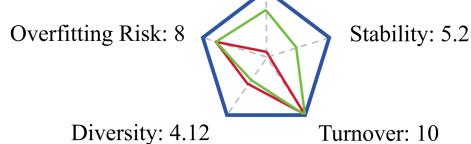
**Node chosen for expansion:**  $V_3$

**The dimension sampled for refinement:** Stability

**Refined Alpha Formula:**  $\text{Tanh}(\text{Sin}(\text{vwap}/\text{Max}(\text{vwap}, 50))) \cdot \text{Tanh}(\text{Skew}(\text{volume}, 45)) \cdot \text{Zscore}(\text{volume}, 30))$

Effectiveness: 7.1

**Evaluation Score:** 6.88



**Refinement history of  $V_4$ :**

$V_0 \rightarrow V_3$ : .....

$V_3 \rightarrow V_4$ :

**Refinement Abstract:** The improved alpha factor enhances the original by incorporating volume skewness to capture crowd-driven activity and a nonlinear regime filter based on vwap extremes, resulting in better detection of regime shifts and reduced sensitivity to short-term noise.

**Score Change:** Effectiveness (+6.8), Stability (+4.8), Turnover (+0), Diversity (-1.63), Overfitting Risk (+0)

**Current / Target Search Num:** 4 / 4

Figure 12: Case Study: expansion to node  $V_4$  and search termination.

When generating alpha formulas, the LLM is instructed to use symbolic parameters (e.g., the lookback window for a moving average operator) rather than fixed numerical values. Concurrently, the LLM proposes several candidate sets of parameter values (three sets in our experiments). We then backtest the alpha derived from each parameter set and select the configuration yielding the best performance. This strategy facilitates a more efficient exploration and utilization of each generated formula structure.

### Refinement Suggestion Generation

During each refinement iteration, once a target dimension for improvement is selected as per Equation 3, we leverage the LLM to generate targeted refinement suggestions. We employ few-shot learning, using alphas from the effective alpha repository  $\mathcal{F}_{zoo}$  as exemplars. The exemplar selection strategy is tailored for each evaluation dimension to provide the most relevant guidance:

**Effectiveness and Stability** To select diverse yet potent exemplars, we first filter the effective alpha repository by removing alphas with the highest correlation to the current alpha (top  $\eta\%$ , where  $\eta = 50$  in our experiments). This creates a candidate pool of more structurally distinct exemplars. From this filtered subset, we select the top- $k$  alphas (e.g.,  $k = 3$ ) exhibiting the highest scores in Effectiveness or Stability to serve as few-shot examples. This two-stage process ensures diversity by preventing highly similar alphas from dominating the exemplar set and encouraging more innovative refinements.

**Diversity** To encourage the exploration of novel alpha structures, we select the top- $k$  alphas from the effective alpha repository  $\mathcal{F}_{zoo}$  that exhibit the lowest correlation with the current alpha as few-shot exemplars.

**Turnover and Overfitting Risk** For these dimensions, we adopt a zero-shot approach, directly prompting the LLM to generate refinement suggestions without explicit exemplars, relying on the model’s intrinsic understanding of these concepts.

Following the generation of a textual suggestion, the LLM produces a revised alpha formula. We then perform a syntax validation on this formula. If it is syntactically incorrect, the LLM receives the error as feedback and is prompted to correct it iteratively until the formula passes validation.

### Multi-Dimensional Evaluation Metrics

As introduced in Section 3, we evaluate each candidate alpha  $f$  across a set of  $q$  dimensions to provide granular feedback for refinement. The score for each dimension,  $e_i$ , is typically derived from its percentile rank against the repository  $\mathcal{F}_{zoo}$  as defined in Equation 7. The primary dimensions used in our framework are detailed as follows:

- Effectiveness: This dimension measures the alpha’s core predictive power.
- Stability: This dimension assesses the consistency of an alpha’s predictive performance over time. An alpha that performs well only in specific market regimes is less reliable.
- Turnover: This dimension evaluates the trading cost associated with an alpha. It is measured by the average daily change in the alpha’s portfolio holdings. A high turnover implies frequent rebalancing, which can incur significant transaction costs and erode profits. The goal is to maintain turnover within a desirable, low range.
- Diversity: This dimension quantifies the novelty an alpha contributes to the existing repository  $\mathcal{F}_{zoo}$ . A valuable alpha should provide predictive signals that are not redundant with those already discovered.
- Overfitting Risk: As described in Section 3, this is a qualitative assessment performed by an LLM. The model analyzes the formula’s complexity, the number of parameters, and its refinement history to identify signs of being overly tailored to the training data (i.e., “p-hacking”), which would impair its generalization performance.

### Dynamic Search Budget Allocation

To balance the breadth and depth of the MCTS search, we employ a dynamic budget allocation strategy. This approach directs more computational resources towards the most promising search paths without prematurely abandoning exploration. The mechanism is governed by two parameters: an initial search budget  $B$  and a budget increment  $b$ .

The search process for any given MCTS tree begins with an initial budget of  $B$  expansion steps. During the search, we monitor the score of each newly generated alpha. If a new alpha  $f_{new}$  achieves a score  $S(f_{new})$  that surpasses the highest score previously seen within that specific MCTS tree, it constitutes a “breakthrough”. Upon each such breakthrough, the remaining search budget for the current tree is increased by the increment  $b$ . The search for the current tree terminates when its allocated budget is exhausted.

This strategy ensures that promising avenues, identified by the discovery of high-quality alphas, are explored more thoroughly, while preventing excessive resource expenditure on less fruitful branches of the search space  $\mathcal{A}$ . It dynamically adapts the search effort based on real-time performance feedback, leading to a more efficient and effective alpha mining process.

## E List of Operators

In this paper, the operators we used can be categorized into two classes: unary operators and binary operators. All operators perform temporal operations, utilizing data from either the current trading day alone or including data from a preceding period. We list the details of all the operators in Table 2

Operator	Type	Description
$-x,  x , x^2, 1/x$	Unary	The opposite/absolute/square/inverse value of $x$ .
$\text{Sign}(x)$		The sign of $x$ .
$\text{Sin}(x), \text{Cos}(x), \text{Tanh}(x)$		Sine, cosine, and hyperbolic tangent functions.
$\text{Log}(x)$		The natural logarithm of $x$ .
$\text{Delay}(x, t)$		The value of expression $x$ at $t$ trading days prior.
$\text{Diff}(x, t)$		The difference between $x$ and the value prior to day $t$ , $x - \text{Delay}(x, t)$ .
$\text{Pct}(x, t)$		The rate of change of $x$ relative to its value $t$ days prior.
$\text{Ma}(x, t), \text{Med}(x, t), \text{Sum}(x, t)$		The mean/median/sum value of $x$ over the past $t$ days.
$\text{Std}(x, t)$		The standard deviation value of $x$ over the past $t$ days.
$\text{Max}(x, t), \text{Min}(x, t)$		The maximum/minimum value of $x$ over the past $t$ days.
$\text{Rank}(x, t)$		The ranking of $x$ relative to its values over the past $t$ days.
$\text{Skew}(x, t), \text{Kurt}(x, t)$		The skewness/kurtosis value of $x$ over the past $t$ days.
$\text{Vari}(x, t)$		The variation value of $x$ over the past $t$ days, $\text{Std}(x, t)/\text{Ma}(x, t)$ .
$\text{Autocorr}(x, t, n)$		The autocorrelation coefficient of $x$ with a lag of $n$ over the past $t$ days.
$\text{Zscore}(x, t)$		The z-score normalization of $x$ based on the mean and standard deviation of its values over the past $t$ days.
$x + y, x - y, x \cdot y, x/y$	Binary	Arithmetic operators.
$\text{Greater}(x, y), \text{Less}(x, y)$		Whether the first value is larger/smaller than the second value.
$\text{Cov}(x, y, t)$		The covariance between time series $x$ and $y$ over the past $t$ days.
$\text{Corr}(x, y, t)$		The Pearson's correlation coefficient between time series $x$ and $y$ over the past $t$ days.

Table 2: All the operators used in the experiments.

## F Pseudo-Code

We provide the pseudo-code for our LLM-guided MCTS framework in Algorithm 1.

## G Experimental Setup Details

### Dataset

All empirical data in this study are obtained through the Qlib platform (Yang et al. 2020), an open-source framework for quantitative financial research. We utilize the daily-frequency data for the Chinese A-shares market as provided by Qlib.

To ensure maximum reproducibility and transparency, we deliberately refrain from applying any additional data filtering, preprocessing, or adjustments. The data is thus used in its original form as sourced from the platform. Furthermore, to maintain a consistent experimental environment, all backtesting simulations are implemented and executed within the Qlib framework. This leverages its integrated backtesting engine and ensures a standardized evaluation process for all experiments.

### Hyperparameter Configurations

**Temperature Parameter of LLMs** When generating alpha portraits and alpha formulas, the temperature is set to 1.0; when correcting illegal alpha formulas, the temperature is adjusted to 0.8; and when scoring the overfitting risk of alpha, the temperature is set to 0.1.

**MCTS** We set the exploration weight  $c$  in the UCT criterion to 1. For each search tree, the initial search budget is 3, and whenever a node achieves a higher score than previously, the search budget is increased by 1.

**Effective Alpha Check** Upon the completion of a search tree's expansion, we examine the effectiveness of the alpha formulas corresponding to all nodes within the tree. Those alphas that pass the effectiveness check are added to the effective alpha repository. The specific criteria for determining effectiveness are as follows:

- Basic Criteria:  $\text{RankIC} \geq 0.015, \text{RankIR} \geq 0.3, R_{\text{RankIC}} \leq 0.95, R_{\text{RankIR}} \leq 0.95$ .
- Turnover Criteria:  $\text{Daily Turnover} \leq 1.6$ .
- Diversity Criteria: The maximum correlation with the alpha within the effective alpha repository is less than 0.8.

After the mining process is completed, we select the top  $k$  alphas with the highest RankIR from the effective alpha repository to form the final alpha set. In our experiments,  $k$  is set to 10, 50, and 100.

---

**Algorithm 1:** LLM-Guided MCTS Framework for Automated Alpha Mining

---

**Input:**  $f_{\text{seed}}$  (initial alpha),  $LLM$  (language model),  $c$  (exploration const.),  $T$  (temperature),  $e_{\max}$  (max score/dim),  $B$  (initial search budget),  $b$  (budget increment),  $\theta_{\text{eff}}$  (effectiveness threshold),  $\mathcal{G}_{\text{forbidden}}$  (initial forbidden structures)

**Output:** Repository  $\mathcal{F}_{\text{zoo}}$  of effective alphas

```

1    $\mathcal{F}_{\text{zoo}} \leftarrow \emptyset$                                 /* Initialization */
2    $s_0 \leftarrow \text{CreateRootNode}(f_{\text{seed}})$ 
3    $\mathbf{E}_{s_0} \leftarrow \text{MultiDimEvaluate}(f_{\text{seed}}, \mathcal{F}_{\text{zoo}})$ 
4    $N_{s_0} \leftarrow 1; S_{s_0} \leftarrow \text{AggregateScore}(\mathbf{E}_{s_0})$  ▷ Store the alpha's intrinsic score  $S(f_{s_0})$ 
5   Tree  $\mathcal{T} \leftarrow \{s_0\}$ 
6    $S_{\max} \leftarrow S_{s_0}$  ▷ Track highest alpha score  $S(f)$  found in the tree
7
8   for  $\text{iter} \leftarrow 1$  to  $B$  do
9        $s_{\text{selected}}, P_{\text{path}} \leftarrow \text{SelectNodeViaUCT}(\mathcal{T}, s_0, c)$  ▷ Select leaf/internal node for expansion.  $P_{\text{path}}$  is a list of  $(s, a)$  pairs.
10      /* 1. Prioritize refinement dimension
11       $\mathbf{E}_s \leftarrow s_{\text{selected}}.\mathbf{E}$ 
12       $P(i|s) \leftarrow \text{Softmax}((e_{\max}\mathbf{1}_q - \mathbf{E}_s)/T)$ 
13       $i^* \leftarrow \text{SampleDimension}(P(i|s))$ 
14      /* 2. LLM generates refinement and new alpha formula
15      context  $\leftarrow \text{GetNodeRefinementContext}(s_{\text{selected}}, \mathcal{T})$  ▷ Parent, children, siblings history
16      examples  $\leftarrow \text{SampleEffectiveAlphas}(\mathcal{F}_{\text{zoo}})$  ▷ Few-shot examples
17       $d_{s,i^*}, f_{\text{new}} \leftarrow LLM.\text{GenerateRefinedAlpha}(s_{\text{selected}}.f, i^*, \text{context}, \text{examples}, \mathcal{G}_{\text{forbidden}})$ 
18      /* 3. Validate and revise formula iteratively
19      while  $\neg \text{IsValid}(f_{\text{new}})$  do
20          feedback  $\leftarrow \text{GetInvalidityReason}(f_{\text{new}})$ 
21           $d_{s,i^*}, f_{\text{new}} \leftarrow LLM.\text{CorrectAlphaFormula}(d_{s,i^*}, \text{feedback}, \text{context}, \text{examples}, \mathcal{G}_{\text{forbidden}})$ 
22      end
23       $\mathbf{E}_{\text{new}} \leftarrow \text{MultiDimEvaluate}(f_{\text{new}}, \mathcal{F}_{\text{zoo}})$  ▷ As in Section 3, using Eq. 6 and LLM for overfitting
24       $S(f_{\text{new}}) \leftarrow \text{AggregateScore}(\mathbf{E}_{\text{new}})$  ▷ Calculate overall score using Eq. 8
25
26      /* Create new node and add to tree
27       $s_{\text{new}} \leftarrow \text{CreateNode}(f_{\text{new}}, \mathbf{E}_{\text{new}}, S(f_{\text{new}}))$ 
28       $\text{AddChildNode}(\mathcal{T}, s_{\text{selected}}, s_{\text{new}})$  ▷ The action leading to  $s_{\text{new}}$  is implicitly defined
29
30      for  $(s_k, a_k)$  in  $P_{\text{path}}$  do
31           $N_{s_k} \leftarrow N_{s_k} + 1$ 
32           $Q(s_k, a_k) \leftarrow \max(Q(s_k, a_k), S(f_{\text{new}}))$  ▷ Update Q-value as per Eq. 10
33      end
34       $N_{s_{\text{new}}} \leftarrow N_{s_{\text{new}}} + 1$ 
35
36      if  $\text{GetEffectivenessScore}(\mathbf{E}_{\text{new}}) \geq \theta_{\text{eff}}$  then
37           $\mathcal{F}_{\text{zoo}} \leftarrow \mathcal{F}_{\text{zoo}} \cup \{f_{\text{new}}\}$ 
38           $\mathcal{G}_{\text{forbidden}} \leftarrow \text{UpdateForbiddenStructures}(\mathcal{F}_{\text{zoo}})$  ▷ Update based on FSA logic
39      end
40      if  $S(f_{\text{new}}) > S_{\max}$  then
41           $B \leftarrow B + b$  ▷ Increase total search budget
42           $S_{\max} \leftarrow S(f_{\text{new}})$  ▷ Update overall max score
43      end
44
45  end
46  return  $\mathcal{F}_{\text{zoo}}$ 

```

---

**Evaluation Score** Here, we present the backtesting metrics utilized to calculate scores for evaluation dimensions: Effectiveness: RankIC, Stability: RankIR, Turnover: Daily turnover rate, Diversity: Maximum correlation with alphas in the effective alpha repository.

**Other Settings** When generating alpha refinement suggestions, the number of few-shot examples is set to 1; in the Frequent Subtree Avoidance method, the number of frequent subtrees to be avoided is set to 3. When calculating the selection probability for each evaluation dimension, the temperature parameter is set to  $T = 1$ . Furthermore, during the generation of refinement suggestions for the Effectiveness and Stability dimensions, the correlation filtering ratio is maintained at  $\eta = 50\%$ .

## Model Settings

To ensure a fair comparison when evaluating alpha pools derived from different methods, we utilize fixed hyperparameters and consistent training strategies for all models.

For the LightGBM model, we configure it with 32 leaves per tree and a total of 200 estimators. The maximum depth of each tree is limited to 8, and the learning rate is set to 0.05. Both L1 and L2 regularization coefficients are fixed at 0.1 to control model complexity. In line with a simple hold-out approach, the model is trained on the entire training dataset (2011/01/01–2020/12/31) for the predetermined number of boosting rounds, without using a separate validation set for early stopping. The fully trained model is then evaluated on the final test set.

Similarly, we train the Multi-Layer Perceptron (MLP) models with a consistent configuration. The MLP architecture consists of three hidden layers with 256, 128, and 64 units, respectively. A dropout rate of 0.3 is applied after each hidden layer to mitigate overfitting. We employ the Adam optimizer with a learning rate of 0.001 and a batch size of 1024, using Mean Squared Error (MSE) as the loss function. To prevent overfitting, early stopping with a patience of 5 epochs is implemented. For this purpose, we partition the data, designating the final year of the training period (2020/01/01–2020/12/31) as the validation set. The model is therefore trained on data from 2011 through 2019, with its performance on the validation set monitored to determine the stopping point.

## Backtesting Strategy

The top- $k$ /drop- $n$  portfolio construction strategy employed in our backtests operates as follows. On each trading day, an equal-weight portfolio is formed by selecting the top  $k$  stocks based on the predictive signals from the trained models. We set  $k$  to represent the top 10% of the respective stock pool (e.g.,  $k = 30$  for the CSI300 index and  $k = 100$  for the CSI1000 index). To manage turnover and limit trading costs, a maximum of  $n$  stocks are bought or sold daily. The value of  $n$  is determined by  $n = k/w$ , where  $w$  is the prediction horizon in days (e.g., for a 10-day return prediction on CSI300,  $n = 30/10 = 3$ ). This strategy ensures that the theoretical complete turnover period of the portfolio aligns with the prediction horizon. A conservative transaction cost of 0.15% per trade is incorporated to ensure a realistic performance assessment.

## Environment

All the experiments are conducted with following settings:

- CPU: AMD EPYC 7642 48-Core Processor
- GPU: NVIDIA GeForce RTX 3080Ti
- Operating system: Ubuntu 20.04.3 LTS
- Software versions: Python 3.8.5; Numpy 1.24.4; Pandas 1.5.2; Pytorch 2.2.2; Openai 1.57.4

## Predictive Performance Evaluation Metrics

To evaluate the predictive performance of the model (trained on the mined alpha set) in forecasting stock returns, we employ several standard metrics: the Information Coefficient (IC), Rank Information Coefficient (RankIC), Annualized Excess Return (AER), and Information Ratio (IR). Let  $f_{i,t}$  denote the predicted returns for asset  $i$  at time  $t$  (out of  $N_t$  assets in the universe at time  $t$ ), and  $r_{i,t+1}$  be its realized total return over the subsequent period (e.g., from  $t$  to  $t + 1$ ). The evaluation spans a total of  $T$  time periods.

**Information Coefficient (IC)** The IC measures the linear correlation between predicted returns and subsequent total returns. It is calculated for each cross-section at time  $t$  ( $IC_t$ ) and then averaged over all  $T$  periods:

$$IC_t = \frac{\sum_{i=1}^{N_t} (f_{i,t} - \bar{f}_t)(r_{i,t+1} - \bar{r}_{t+1})}{\sqrt{\sum_{i=1}^{N_t} (f_{i,t} - \bar{f}_t)^2} \sqrt{\sum_{i=1}^{N_t} (r_{i,t+1} - \bar{r}_{t+1})^2}} \quad (13)$$

$$IC = \frac{1}{T} \sum_{t=1}^T IC_t \quad (14)$$

where  $\bar{f}_t$  and  $\bar{r}_{t+1}$  are the cross-sectional means of predicted signals and realized total returns, respectively. A higher IC indicates better predictive power.

Instruments	$\Delta T$	Alpha Set	LightGBM				MLP			
			IC	RankIC	AER	IR	IC	RankIC	AER	IR
CSI300	10	Alpha158	0.0386	0.0377	0.0762	0.6713	0.0337	0.0329	0.0307	0.2929
		Alpha360	0.0061	-0.0096	-0.0032	-0.0241	0.0153	0.0231	0.0378	0.2847
		AlphaAgent	0.0298	0.0244	0.0383	0.3319	0.0386	0.0336	0.0437	0.4411
		RiskMiner	0.0412	0.0356	-0.0390	-0.5020	0.0414	0.0405	0.0429	0.5726
		Ours	<b>0.0420</b>	<b>0.0395</b>	<b>0.0822</b>	<b>0.9397</b>	<b>0.0422</b>	<b>0.0408</b>	<b>0.0737</b>	<b>0.8103</b>
	30	Alpha158	0.0028	0.0103	-0.0508	-0.3716	0.0036	0.0036	-0.0122	-0.1083
		Alpha360	-0.0539	-0.0702	-0.0992	-0.5688	0.0235	0.0340	-0.0136	-0.1553
		AlphaAgent	0.0372	0.0339	0.0412	0.3714	0.0363	0.0344	0.0615	0.7084
		RiskMiner	<b>0.0501</b>	<b>0.0453</b>	0.0104	0.1151	0.0407	0.0395	-0.0332	-0.3832
		Ours	0.0417	0.0401	<b>0.0826</b>	<b>1.0312</b>	<b>0.0423</b>	<b>0.0424</b>	<b>0.1315</b>	<b>1.4307</b>
CSI1000	10	Alpha158	0.0610	0.0627	0.0316	0.3919	0.0520	0.0567	0.5877	1.5969
		Alpha360	0.0636	0.0642	0.0798	0.9242	0.0550	0.0608	0.0689	0.7576
		AlphaAgent	0.0726	0.0718	0.0759	0.7111	<b>0.0752</b>	<b>0.0659</b>	0.1087	1.2615
		RiskMiner	0.0752	0.0708	0.0895	0.8997	0.0712	0.0635	0.0709	0.7872
		Ours	<b>0.0804</b>	<b>0.0729</b>	<b>0.1393</b>	<b>1.3577</b>	0.0662	0.0618	<b>0.1204</b>	<b>1.2868</b>
	30	Alpha158	0.0490	0.0654	0.0343	0.5068	0.0380	0.0373	0.0545	0.7720
		Alpha360	0.0420	0.0567	0.0171	0.2692	0.0532	0.0456	0.0887	1.2959
		AlphaAgent	0.0778	0.0692	0.0756	0.9194	0.0713	0.0660	0.0871	1.0454
		RiskMiner	0.0701	0.0686	0.1022	1.1738	0.0722	0.0651	0.0843	1.0918
		Ours	<b>0.0793</b>	<b>0.0723</b>	<b>0.1326</b>	<b>1.2598</b>	<b>0.0738</b>	<b>0.0710</b>	<b>0.1696</b>	<b>1.5695</b>

Table 3: Predictive performance comparison between our framework and other baselines. For AlphaAgent, RiskMiner and our framework, the size of alpha sets is fixed at 100.

**Rank Information Coefficient (RankIC)** The RankIC measures the monotonic relationship (Spearman’s rank correlation) between predicted returns and subsequent total returns. It is less sensitive to outliers than the Pearson correlation-based IC. Similar to IC, it is computed cross-sectionally and then averaged:

$$\text{RankIC}_t = \text{Corr}(\text{rank}(f_{1,t}, \dots, f_{N_t,t}), \text{rank}(r_{1,t+1}, \dots, r_{N_t,t+1})) \quad (15)$$

$$\text{RankIC} = \frac{1}{T} \sum_{t=1}^T \text{RankIC}_t \quad (16)$$

where  $\text{rank}(\cdot)$  denotes the operation that assigns ranks to the elements in the vector, and  $\text{Corr}(\cdot, \cdot)$  is the Pearson correlation coefficient applied to these ranks.

**Annualized Excess Return (AER)** The AER measures the simple arithmetic average rate of excess return per year generated by a portfolio against a market benchmark. For our long-only strategy, at each rebalancing period  $t$ , we select the top- $k$  assets with the highest predicted returns  $f_{i,t}$ . The portfolio’s performance is based on the realized excess returns of these assets. Let  $r_{i,t+1}^e$  denote the realized excess return of asset  $i$  over the market benchmark for the period from  $t$  to  $t + 1$ . Assuming the  $k$  selected assets are equally weighted, the portfolio’s excess return for the period  $t + 1$ , denoted  $R_{p,t+1}$ , is the average of the individual assets’ excess returns:

$$R_{p,t+1} = \frac{1}{k} \sum_{s \in \text{TopK}_t} r_{s,t+1}^e \quad (17)$$

where  $\text{TopK}_t$  is the set of  $k$  assets with the highest predicted returns based on  $f_t$ . The AER, calculated as the arithmetic mean of these per-period portfolio excess returns scaled to an annual figure, over  $T_p$  portfolio holding periods is then:

$$\text{AER} = \left( \frac{1}{T_p} \sum_{j=1}^{T_p} R_{p,j} \right) \times P \quad (18)$$

where  $R_{p,j}$  is the portfolio excess return over the benchmark in the  $j$ -th holding period,  $T_p$  is the total number of holding periods in the backtest, and  $P$  is the number of holding periods in a year (e.g.,  $P = 252$  for daily rebalancing,  $P = 12$  for monthly rebalancing).

Model	Method	Alpha Num = 10		Alpha Num = 50		Alpha Num = 100	
		IC	RankIC	IC	RankIC	IC	RankIC
LightGBM	GP	-0.0055	-0.0055	0.0032	0.0039	0.0060	0.0072
	DSO	0.0022	0.0005	0.0038	0.0044	0.0006	0.0015
	AlphaGen	0.0094	<b>0.0106</b>	0.0125	0.0121	0.0126	<b>0.0134</b>
	AlphaForge	0.0016	0.0007	0.0086	0.0060	0.0110	0.0094
	CoT	0.0014	0.0002	0.0106	0.0104	0.0116	0.0112
	ToT	0.0032	0.0029	0.0098	0.0111	0.0102	0.0103
	FAMA	0.0086	0.0080	0.0085	0.0095	0.0119	0.0120
Ours		<b>0.0097</b>	0.0102	<b>0.0128</b>	<b>0.0129</b>	<b>0.0132</b>	0.0130
MLP	GP	-0.0063	-0.0070	0.0048	0.0051	0.0041	0.0047
	DSO	-0.0017	-0.0018	0.0080	0.0079	0.0031	0.0036
	AlphaGen	0.0089	0.0094	0.0124	0.0118	0.0117	0.0112
	AlphaForge	-0.0001	0.0000	0.0103	0.0104	0.0070	0.0073
	CoT	0.0029	0.0035	0.0115	0.0116	0.0113	0.0114
	ToT	0.0029	0.0029	0.0092	0.0096	0.0120	0.0113
	FAMA	0.0079	0.0080	0.0109	0.0112	0.0111	0.0108
Ours		<b>0.0092</b>	<b>0.0097</b>	<b>0.0130</b>	<b>0.0127</b>	<b>0.0125</b>	<b>0.0122</b>

Table 4: The predictive performance alphas mined by different methods on the S&P500 stock pool.

**Information Ratio (IR)** The IR measures the risk-adjusted excess return of the portfolio. It is defined as the Annualized Excess Return divided by its annualized volatility:

$$\text{IR} = \frac{\text{AER}}{\sigma(R_p)\sqrt{P}} \quad (19)$$

where AER is the Annualized Excess Return calculated as per Eq. (18), and  $\sigma(R_p)$  is the standard deviation of the *per-period* portfolio excess returns  $R_{p,j}$  over the  $T_p$  holding periods. A higher IR indicates better return per unit of risk.

## H Additional Results

### Comparisons with Other Baselines

To provide a comprehensive assessment of our proposed framework, we conduct additional comparisons against four benchmark alpha sets. The first two are the widely adopted handcrafted libraries **Alpha158** and **Alpha360**. Alpha158 consists of 158 factors derived from historical price and volume data for the Chinese A-share market, while Alpha360 is a more extensive set of 360 alphas. We construct these benchmark sets based on the information provided on the Qlib platform (Yang et al. 2020). As a third baseline, we include factors generated by **AlphaAgent** (Tang et al. 2025), a modern LLM-driven approach adapted from the RD-Agent (Yang et al. 2025). In contrast to purely search-based methods, AlphaAgent leverages human knowledge provided via natural language prompts to guide its discovery process. For our experiments, we direct it to generate 100 factors using the conceptual prompt of discovering “high-quality alpha factors for the Chinese A-share market.” As a final baseline, we incorporate **RiskMiner** (Ren et al. 2024), a state-of-the-art alpha mining method based on a risk-seeking Monte Carlo Tree Search.

We evaluate the predictive performance of two distinct models: LightGBM and a 3-layer Multi-Layer Perceptron (MLP). Each model is trained using features from five separate sets: (i) Alpha158, (ii) Alpha360, (iii) the AlphaAgent-generated factors, (iv) the RiskMiner-generated factors, and (v) the alphas mined by our proposed method. The comparative results are presented in Table 3. As illustrated, our method consistently achieves superior predictive performance across the various experimental settings and model architectures considered, underscoring its effectiveness in identifying more potent alpha signals compared to these established benchmarks.

### Additional Results on the U.S. Stock Market

To further assess the generalizability and robustness of our proposed framework, we conduct additional experiments on the U.S. stock market, specifically focusing on the S&P 500 index constituents. This allows us to validate whether the effectiveness of our alpha mining approach extends to different market environments.

**Experimental Setup** We utilize the daily U.S. stock market data provided by the Qlib platform (Yang et al. 2020), which includes five fundamental features: open, high, low, close prices, and trading volume. The stock pool consists of the constituents of the S&P 500 index. The prediction target is the 10-day forward return. Following a chronological split, the dataset is divided

LLM	IC	RankIC	IR	RankIR
Random	0.0126	0.0179	0.095	0.148
GPT4.1	0.0130	0.0242	0.097	0.168
Gemini2.0-flash-lite	0.0125	0.0231	0.094	0.180
Deepseek-v3-0324	0.0122	0.0223	0.087	0.145
Ours	<b>0.0527</b>	<b>0.0714</b>	<b>0.368</b>	<b>0.467</b>

Table 5: Investigation of pre-existing knowledge (data leakage) in LLMs. We compare the performance of alphas generated by directly prompting various LLMs for high-performing formulas against a random baseline and our MCTS-based method. Metrics (IC, RankIC, IR, RankIR) are averages over 10 alphas per method. LLMs (GPT4.1, Gemini2.0-flash-lite, Deepseek-v3) are prompted for high-performing alphas (CSI300, 10-day return). ‘Ours’ is our method with an alpha set size of 10.

into a training period from 2007/01/01 to 2015/12/31, and a testing period from 2016/01/01 to 2020/10/10. All other experimental settings, including the alpha generation methods, model configurations (LightGBM and MLP), and evaluation metrics (IC and RankIC), remain consistent with those used for the Chinese market experiments described in the main paper.

**Results** Table 4 presents the predictive performance of alphas generated by our method and the baselines on the S&P 500 dataset. The results demonstrate that our proposed framework maintains a strong competitive advantage in the U.S. market. Across both LightGBM and MLP models and for all alpha set sizes (10, 50, and 100), our method consistently achieves the highest or near-highest IC and RankIC scores. This suggests that the alphas mined by our framework are not only effective in the Chinese market but also possess robust predictive power in the more mature and competitive U.S. market, confirming the broad applicability of our approach.

### Investigation of Potential Data Leakage in LLMs

A critical concern when employing LLMs for tasks like alpha mining is the potential for data leakage. This refers to the possibility that LLMs’ training data might have inadvertently included information about historically well-performing alpha formulas, which could lead to an overestimation of their generative capabilities if they simply recall these formulas.

To investigate this, we design an experiment to explicitly probe whether pre-trained LLMs possess such inherent knowledge. We prompt three distinct LLM backbones—GPT-4.1, Gemini2.0-flash-lite, and Deepseek-v3—to generate alpha formulas specifically anticipated to yield high performance for 10-day forward return prediction on the A-share CSI300 stock pool. For each LLM and the Random baseline, we generate 10 distinct alpha formulas, and the performance metrics reported are averages across these 10 alphas. This setup directly attempts to leverage any pre-existing, “leaked” knowledge within the LLMs. The performance of alphas generated under this explicit instruction is compared against two baselines: (1) alphas generated randomly by an LLM (“Random”), and (2) alphas mined by our proposed framework (“Ours”).

As shown in Table 5, the average performance of alphas from LLMs explicitly prompted to generate high-performing expressions is not significantly different from the Random baseline. All these direct LLM generation approaches yield substantially lower performance across all metrics compared to our framework. These results suggest that the tested LLMs do not inherently leverage leaked knowledge of superior alpha formulas for this task, thereby alleviating data leakage concerns and highlighting the efficacy of our alpha mining framework.

### Sensitivity Analysis of LLM Backbone Choice

We evaluate the impact of different LLM backbones on our framework. We test several models from different series: GPT-4.1, Gemini-2.0-flash-lite, Gemini-2.5-flash-lite-preview-06-17, Deepseek-v3, and Qwen3-235b-a22b-2507. Experiments are conducted on the CSI300 stock pool, predicting 10-day forward returns with an alpha set size of 50. Results are detailed in Table 6. The findings reveal that the choice of LLM backbone leads to notable variations in the performance characteristics of the generated alpha sets. While each LLM may emphasize different aspects of alpha quality, our framework, when leveraging these advanced models, generally demonstrates performance that is competitive with or surpasses the best results of baselines.

### Characteristics of Alphas at Varying MCTS Search Depths

In this section, we analysis the characteristics of alphas derived from nodes at different depths within the MCTS search tree. Alphas located at deeper levels of the tree undergo a greater number of refinement iterations. Figure 14a illustrates how several key metrics evolve with MCTS depth: specifically, the average formula depth (depth of the formula’s tree representation), formula length (number of operators), in-sample (IS) and out-of-sample (OOS) RankIC, and the Overfitting Risk Score.

We observe that as MCTS search depth increases, both the average formula depth and length initially rise before tending to stabilize. This stabilization is largely attributed to the Overfitting Risk Score, which penalizes overly complex formulas more prone to overfitting. Concurrently, both average IS and OOS RankIC exhibit a general upward trend with increasing depth, reflecting the efficacy of the refinement process. Conversely, the generalization gap (the difference between IS and OOS RankIC)

LLM	LightGBM				MLP			
	IC	RankIC	AR	IR	IC	RankIC	AR	IR
Best of Baselines	0.0388	0.0378	0.0680	0.7470	0.0380	0.0374	0.0805	0.8414
GPT4.1	<b>0.0399</b>	<b>0.0394</b>	0.0717	0.8283	<b>0.0436</b>	<b>0.0425</b>	0.0661	0.7075
Gemini-2.0-flash-lite	0.0376	0.0363	0.1037	1.1096	0.0389	0.0372	<b>0.1734</b>	<b>1.9801</b>
Gemini-2.5-flash-lite-preview-06-17	0.0396	0.0390	<b>0.1075</b>	1.1885	0.0375	0.0373	0.1225	1.3898
Deepseek-v3	0.0388	0.0386	0.1064	<b>1.2144</b>	0.0409	0.0401	0.1202	1.3732
Qwen3-235b-a22b-2507	0.0382	0.0372	0.0614	0.7769	0.0386	0.0380	0.1101	1.3177

Table 6: Performance comparison of our framework using different LLM backbones on the CSI300 dataset (10-day forward returns, alpha set size of 50). Best results are highlighted in **bold**.

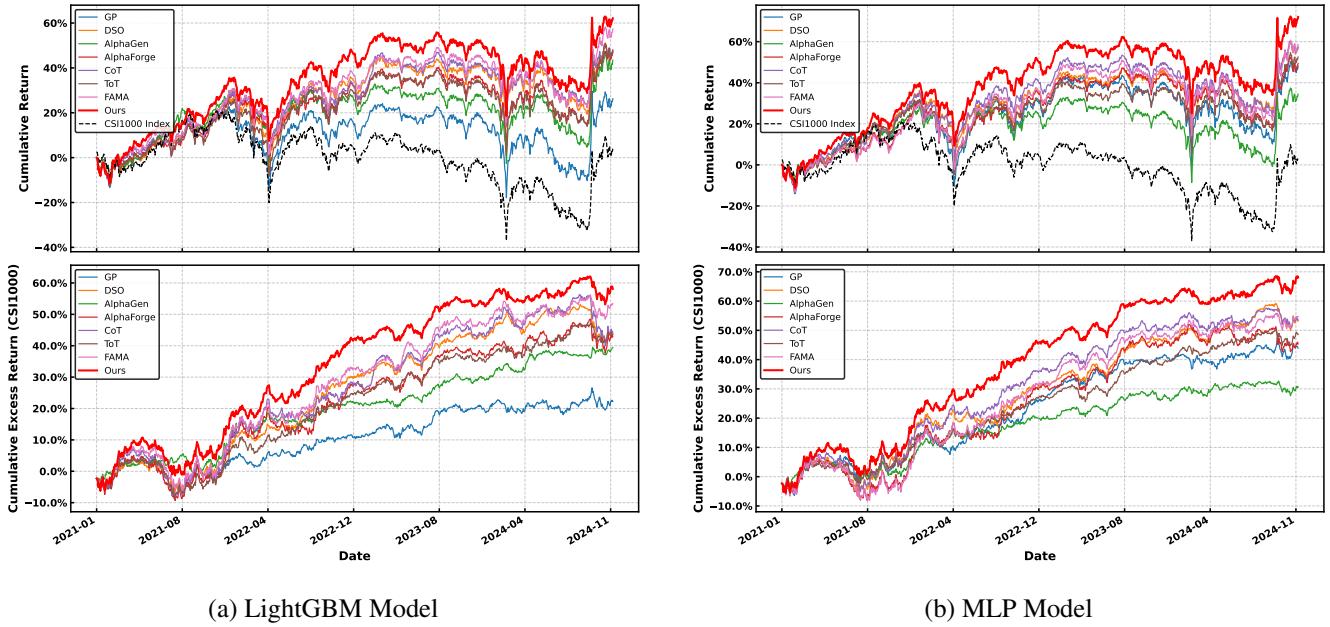


Figure 13: Cumulative return curves of backtest using alphas generated by different methods.

tends to widen, indicating a decline in generalization performance as the formulas become more complex. This trend aligns with the behavior of the Overfitting Risk Score, suggesting that while refinement enhances performance, the associated risk of overfitting necessitates careful management, a role our ORS aims to fulfill.

### Cumulative Return Curve Visualizations

Figure 13 presents the cumulative return curves derived from backtesting using alphas mined by different methods. As illustrated, our proposed method consistently demonstrates superior performance, achieving the highest cumulative returns among the evaluated methods.

### Interpretability of Mined Alpha Formulas

In quantitative finance, interpretability refers to the ability to connect a mathematical factor to a coherent economic rationale explaining why it might predict cross-sectional returns. An interpretable alpha factor should be grounded in established financial theories or well-documented market phenomena, such as momentum, reversal, liquidity, or volatility effects (e.g., (Efficiency 1993; De Bondt and Thaler 1985; Amihud and Mendelson 1986; French, Schwert, and Stambaugh 1987)). Such grounding is valuable as it provides a plausible economic hypothesis for a factor's potential efficacy. This enhances confidence that the factor captures a genuine market dynamic rather than a spurious, overfitted relationship unlikely to generalize out-of-sample.

As shown in Table 7, the formulas discovered by our framework can be deconstructed into components with plausible economic intuition. For instance, the formula  $Zscore(Ma(close - vwap, 20), 30)$  can be interpreted as a measure of anomalous intraday momentum. It isolates the end-of-day buying or selling pressure ( $close - vwap$ ), smooths it into a short-term trend, and then

Method	Alpha Formula
Ours	<ol style="list-style-type: none"> <li>1. Zscore(Ma(close – vwap, 20), 30)</li> <li>2. Std(Pct(vwap, 20), 25) · Sum(volume, 40)/volume</li> <li>3. Corr(close, volume, 50) · Zscore(Ma(close – vwap, 30), 40)</li> <li>4. Diff(Ma(volume, 20), 3)/Ma(volume, 60)</li> <li>5. Corr(Pct(close, 10), Pct(volume, 10), 10) · Corr(Pct(close, 30), Pct(volume, 30), 30) · Skew(volume, 20)</li> <li>6. Ma(Corr(volume, close, 20) · Skew(high – low, 20), 10)</li> </ol>
GP	<ol style="list-style-type: none"> <li>1. Add(Mul(–0.01, volume), Log(Log(close)))</li> <li>2. Less(Cov(open, Add(high, Div(volume, –5)), 10), Std(Log(close), 50))</li> </ol>
DSO	<ol style="list-style-type: none"> <li>1. Greater(volume, Med(Sub(Ma(open, 10), Med(Std(Sign(close), 10), 10)), 10))</li> <li>2. Cov(Med(Sign(vwap), 50), 5, 20)</li> </ol>
AlphaGen	<ol style="list-style-type: none"> <li>1. Corr(Rank(Diff(Greater(2.0, volume), 50), 10), close, 20)</li> <li>2. Ma(Greater(Std(Less(0.01, Less(Div(Log(high), –2.0), –30)), 1), –30), 5)</li> </ol>
AlphaForge	<ol style="list-style-type: none"> <li>1. <math>1/(1/(Diff(\text{Sin}(1/(\text{Cos}(((0.01 + \text{Sin}(\text{Tanh}(\text{high}))))/30)))), 30)))</math></li> <li><math>( (\text{Cos} (\text{Tanh}(\text{Sin}(\text{Diff}(\text{Sin}(30 + \text{low}), 20)))) ) + 30.0 ^2</math></li> </ol>

Table 7: Examples of alpha formulas mined by our framework and other non-LLM-based baselines. Each numbered item represents a distinct alpha formula.

uses a Zscore to flag statistically significant deviations. Such abnormal pressure may signal informed trading or strong sentiment, potentially predicting continued returns. Another example,  $\text{Std}(\text{Pct}(\text{vwap}, 20), 25) \cdot \text{Sum}(\text{volume}, 40)/\text{volume}$ , combines a measure of recent price volatility with a volume-based reversal signal. This formula identifies assets with high recent volatility but abnormally low current trading activity—a potential signature of trend exhaustion that often precedes price reversals. Other formulas generated by our framework are similarly composed of interpretable concepts like price-volume correlation and volume acceleration (see Table 7, formulas 3-6).

In stark contrast, formulas from the baselines often exhibit fundamental issues that undermine their economic plausibility. A primary concern is dimensional inconsistency. For example, the GP formula  $\text{Add}(\text{Mul}(–0.01, \text{volume}), \text{Log}(\text{Log}(\text{close})))$  attempts to add a term based on trading volume (unit: shares) to a transformed price term (unit: log-currency). This is analogous to adding mass to length, an operation that lacks a coherent real-world interpretation. Other methods produce expressions like  $\text{Std}(\text{Less}(...))$ , which calculates the standard deviation of a binary true/false series, a statistically questionable operation. The challenge is most acute for AlphaForge, whose unconstrained search yields formulas like  $1/(1/(Diff(\text{Sin}(1/(\text{Cos}(...))))))$ . The nested trigonometric functions applied to financial data are mathematically valid but have no clear basis in economic theory.

While such formulas might fit the training data, their lack of a plausible economic basis makes them resemble “black boxes.” This is a critical drawback, as formulas that violate basic principles like dimensional consistency are at high risk of being spurious artifacts of overfitting. Consequently, they are less likely to be trusted by practitioners and integrated into investment strategies.

### Sensitivity Analysis of Key Framework Hyperparameters

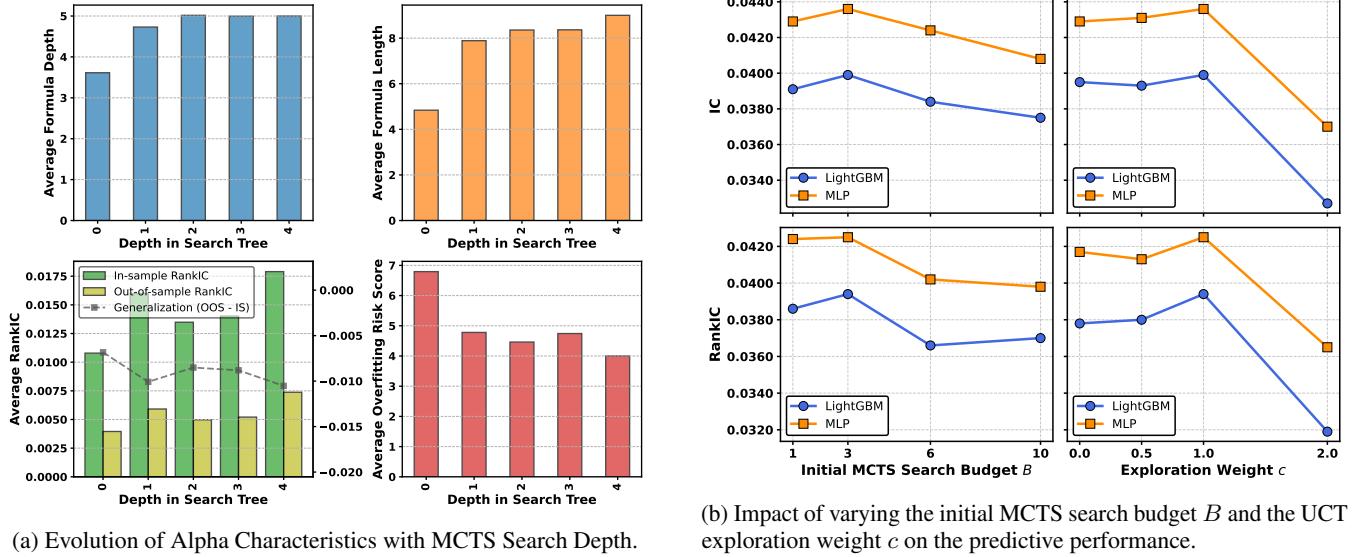
In this section, we investigate the sensitivity of our framework to two key hyperparameters: the initial MCTS search budget  $B$  (i.e., the initial target search number of MCTS) and the UCT exploration weight  $c$ . We conduct experiments using constituents of the CSI300 Index as the stock pool, targeting 10-day forward returns. The alpha set size is set to 50. Figure 14b illustrates the model’s performance under various settings for these parameters.

For the initial MCTS search budget  $B$ , a value of 3 yields the best performance. A larger  $B$  can lead to excessive exploration of unpromising MCTS trees, consuming more computational resources without proportional gains. Conversely, a budget that is too small may result in insufficient exploration of potentially promising trees, thereby diminishing search efficacy.

Regarding the UCT exploration weight  $c$ , our findings suggest that an excessively large value is detrimental. A higher  $c$  biases the search towards more uniform exploration (randomness), potentially squandering refinement opportunities on underperforming alpha candidates.

### Cost Estimation Details

To provide a standardized and equitable basis for comparing the efficiency of different alpha discovery methods, we conduct a cost-performance analysis. This appendix details the methodology we use to estimate the monetary cost of a single experimental



(a) Evolution of Alpha Characteristics with MCTS Search Depth.

(b) Impact of varying the initial MCTS search budget  $B$  and the UCT exploration weight  $c$  on the predictive performance.

Figure 14: Analysis of MCTS search characteristics and parameter sensitivity.

Method	Time (h)	GPU	Input Tokens	Output Tokens	Cost (\$)		
					Server	API	Total
GP	28.30	✓	0	0	87.051	0.000	87.051
DSO	8.70	✓	0	0	26.761	0.000	26.761
AlphaGen	13.80	✓	0	0	42.449	0.000	42.449
AlphaForge	0.42	✓	0	0	1.292	0.000	1.292
CoT	1.42	✗	6,110,411	1,143,407	2.939	21.368	24.307
ToT	1.73	✗	7,072,398	1,800,932	3.581	28.552	32.133
FAMA	1.81	✗	6,866,434	1,198,739	3.747	23.323	27.069
Ours (GPT4.1)	2.76	✗	21,391,823	3,238,154	5.713	68.689	74.402
Ours (Gemini-2.0-flash-lite)	1.94	✗	24,175,761	5,692,047	4.016	3.521	7.537
Ours (Deepseek-v3)	6.91	✗	27,291,214	5,257,428	14.304	7.931	22.235

Table 8: Comparison of single-experiment costs and resource usage across different methods. Time is measured in hours. GPU usage is indicated by a checkmark (✓) for required or a cross (✗) for not required. Costs are presented in USD.

run for each method, as presented in Table 8. The cost estimation is based on a representative experimental setting: generating an alpha set of 100 alphas on the CSI 300 index.

The total cost for a single run comprises two primary components: server computational cost and LLM API cost.

**Server Computational Cost** The server cost is a function of the experiment’s runtime and the required hardware. We reference public cloud computing prices from Amazon Web Services (AWS) for our calculations.

- CPU-only Server:** For methods that do not require a GPU (all LLM-based methods), we use a 48-core CPU server, priced at \$2.07 per hour.
- GPU-enabled Server:** For methods requiring GPU acceleration (all non-LLM-based methods), we use a server equipped with an RTX 3080Ti GPU. The total server cost combines the base CPU server price with the GPU price of \$1.006 per hour, for a total hourly rate of \$3.076.

The server cost is calculated as:  $\text{Cost}_{\text{Server}} = \text{Runtime (h)} \times \text{Hourly Rate (\$/h)}$ .

**LLM API Cost** The API cost applies only to LLM-based methods and depends on the number of input and output tokens processed. We use the official pricing for each model. The costs, specified in USD per million tokens for input and output respectively, are as follows:

- GPT-4.1:** (\$2.00, \$8.00)
- Gemini-2.0-flash-lite:** (\$0.075, \$0.30)

- **Deepseek-v3:** (\$0.071, \$1.14)

The API cost is calculated as:  $\text{Cost}_{\text{API}} = \left( \frac{\text{Input Tokens}}{10^6} \times \text{Rate}_{\text{Input}} \right) + \left( \frac{\text{Output Tokens}}{10^6} \times \text{Rate}_{\text{Output}} \right)$ .

**Estimation Considerations** It is important to acknowledge two caveats in our estimation:

- The cost for non-LLM-based methods is likely an underestimation. Real-world quantitative finance applications often involve higher-frequency data, which would significantly increase the time and computational resources for alpha factor calculation and backtesting.
- Conversely, the API cost for LLM-based methods may be a slight overestimation, as we do not account for potential cost reductions from API-level caching, where repeated input tokens can be priced lower.

The final total cost in Table 8 is the sum of the server and API costs, offering a unified metric to evaluate search efficiency in terms of performance per dollar.

Model	Method	Alpha Num = 10				Alpha Num = 50				Alpha Num = 100			
		IC	RankIC	AR	IR	IC	RankIC	AR	IR	IC	RankIC	AR	IR
LightGBM	CoT	0.0201	0.0206	0.0549	0.6713	0.0314	0.0289	0.0531	0.5939	0.0373	0.0352	0.0321	0.3992
	ToT	0.0269	0.0267	0.0387	0.4335	0.0261	0.0240	0.0695	0.7544	0.0358	0.0332	0.0620	0.6872
	FAMA	0.0210	0.0206	0.0175	0.2103	0.0243	0.0227	0.0593	0.7470	0.0366	0.0345	0.0731	0.8009
	Ours	<b>0.0386</b>	<b>0.0364</b>	<b>0.0668</b>	<b>0.7485</b>	<b>0.0399</b>	<b>0.0394</b>	<b>0.0717</b>	<b>0.8283</b>	<b>0.0420</b>	<b>0.0395</b>	<b>0.0822</b>	<b>0.9397</b>
MLP	CoT	0.0207	0.0209	0.0608	0.7030	0.0304	0.0290	0.0566	0.6779	0.0367	0.0334	0.0343	0.4163
	ToT	0.0281	0.0278	0.0625	0.7185	0.0299	0.0291	0.0571	0.6340	0.0323	0.0316	0.0340	0.3770
	FAMA	0.0243	0.0227	0.0593	0.7470	0.0292	0.0281	0.0585	0.6868	0.0378	0.0373	<b>0.0845</b>	<b>0.9401</b>
	Ours	<b>0.0411</b>	<b>0.0406</b>	<b>0.0741</b>	<b>0.8186</b>	<b>0.0436</b>	<b>0.0425</b>	<b>0.0661</b>	<b>0.7075</b>	<b>0.0422</b>	<b>0.0408</b>	0.0737	0.8103

Table 9: Predictive performance of LightGBM and MLP models trained on alphas mined by different methods. The experiment is conducted on the CSI 300 stock pool with a 10-day return prediction horizon.

**Comparison under Equal API Cost** To further scrutinize the search efficiency of the LLM-based methods, we conduct an additional comparison under an equal-cost constraint. For the baseline methods, we incrementally increase the number of generated alphas, evaluating performance at each 1,000-alpha interval until their total API cost matches that of our method. We report the results from the evaluation interval that achieved the highest IC. As shown in Table 9, our method still achieves superior performance even when baselines are allocated an equivalent API budget. To match this budget, the baseline methods require generating a larger volume of alphas, and their runtimes exceed that of our method. This analysis suggests that our framework demonstrates high search efficiency across the dimensions of search count, token consumption, and runtime.

## I Full Experimental Results

In this section, we provide detailed experimental results for our proposed method and all other baselines. Table 10 and 11 illustrate the experimental results for the LightGBM and MLP models on the CSI300 stock pool. Similarly, Table 12 and 13 present the results for these two models on the CSI1000 stock pool.

## J Limitations

Despite the promising advancements presented by our framework for formulaic alpha mining, several limitations warrant discussion.

First, while our method generates effective alpha formulas, there remains a gap in achieving the same level of novelty and complexity typically found in formulas developed by human experts. The framework may sometimes struggle to produce highly intricate or unconventional alphas. Second, the diversity of the generated alphas is inherently constrained by the internal knowledge base of LLMs. This can, in turn, limit the breadth of the search space explored compared to certain non-LLM-based methodologies. Consequently, scaling our approach to extremely large-scale alpha mining tasks, which necessitate the exploration of a vast and diverse alpha landscape, may present challenges.

Addressing these limitations, such as by incorporating mechanisms for enhanced novelty or expanding the effective search space, constitutes important directions for future research.

## K LLM Agent Prompts

In our proposed framework, LLMs function as autonomous agents, instrumental in the multifaceted process of alpha formula mining. This includes the generation, iterative refinement, and overfitting risk evaluation of alpha formulas. This section details the core prompts designed to guide the LLM in these pivotal tasks. Each prompt is carefully designed to elicit specific behaviors and outputs from the LLM, ensuring a structured and effective approach to mining alpha formulas.

$\Delta T$	Method	Alpha Num = 10				Alpha Num = 50				Alpha Num = 100			
		IC	RankIC	AR	IR	IC	RankIC	AR	IR	IC	RankIC	AR	IR
10	GP	0.0165	0.0148	0.0138	0.1717	0.0248	0.0246	0.0570	0.6777	0.0319	0.0288	0.0716	0.7222
	DSO	0.0079	0.0090	0.0434	0.6125	0.0174	0.0178	-0.0569	-0.6460	0.0246	0.0247	-0.0090	-0.1069
	AlphaGen	0.0443	0.0411	0.0119	0.1335	0.0388	0.0378	-0.0342	-0.3473	<b>0.0446</b>	<b>0.0416</b>	0.0094	0.1091
	AlphaForge	<b>0.0521</b>	<b>0.0518</b>	0.0231	0.2653	0.0293	0.0271	0.0252	0.3005	0.0407	0.0387	-0.0253	-0.3004
	CoT	0.0201	0.0206	0.0549	0.6713	0.0240	0.0224	0.0600	0.6683	0.0237	0.0225	0.0600	0.6681
	ToT	0.0269	0.0267	0.0387	0.4335	0.0256	0.0242	0.0680	0.7365	0.0358	0.0332	0.0620	0.6872
	FAMA	0.0210	0.0206	0.0175	0.2103	0.0243	0.0227	0.0593	0.7470	0.0292	0.0281	0.0585	0.6868
Ours		0.0386	0.0364	<b>0.0668</b>	<b>0.7485</b>	<b>0.0399</b>	<b>0.0394</b>	<b>0.0717</b>	<b>0.8283</b>	0.0420	0.0395	<b>0.0822</b>	<b>0.9397</b>
30	GP	0.0145	0.0118	0.0322	0.3857	0.0248	0.0246	0.0570	0.6777	0.0319	0.0288	0.0716	0.7222
	DSO	0.0079	0.0090	-0.0050	-0.0682	0.0174	0.0178	0.0156	0.1735	0.0246	0.0247	0.0185	0.2039
	AlphaGen	0.0299	0.0280	0.0597	0.7792	0.0322	0.0286	0.0150	0.1540	0.0401	0.0339	0.0548	0.4662
	AlphaForge	0.0326	0.0322	0.0981	0.9455	0.0286	0.0278	0.0551	0.7341	0.0339	0.0315	0.0164	0.2176
	CoT	0.0200	0.0186	0.0774	0.9554	0.0274	0.0250	0.0574	0.6105	0.0278	0.0247	0.0468	0.5037
	ToT	0.0232	0.0247	0.1064	1.3221	0.0291	0.0285	0.0632	0.7982	0.0348	0.0312	0.0587	0.7037
	FAMA	0.0298	0.0307	0.1050	1.1609	0.0285	0.0298	0.0803	0.9502	0.0301	0.0297	<b>0.0934</b>	<b>1.1629</b>
Ours		<b>0.0334</b>	<b>0.0334</b>	<b>0.1129</b>	<b>1.3286</b>	<b>0.0352</b>	<b>0.0340</b>	<b>0.0886</b>	<b>1.1299</b>	<b>0.0417</b>	<b>0.0401</b>	0.0826	1.0312

Table 10: The predictive performance of LightGBM model trained on alphas mined by different methods on the CSI300 stock pool.

$\Delta T$	Method	Alpha Num = 10				Alpha Num = 50				Alpha Num = 100			
		IC	RankIC	AR	IR	IC	RankIC	AR	IR	IC	RankIC	AR	IR
10	GP	0.0124	0.0120	-0.0212	-0.2240	0.0257	0.0233	0.0769	0.8324	0.0277	0.0266	0.0345	0.3967
	DSO	0.0085	0.0082	0.0485	0.5849	0.0062	0.0052	-0.0263	-0.2623	0.0256	0.0245	0.0019	0.0230
	AlphaGen	0.0419	0.0418	0.0725	0.8199	0.0380	0.0374	0.0524	0.6187	0.0311	0.0313	0.0295	0.3044
	AlphaForge	<b>0.0549</b>	<b>0.0554</b>	<b>0.0802</b>	<b>0.8397</b>	0.0232	0.0207	0.0668	0.7509	0.0408	0.0395	0.0220	0.2555
	CoT	0.0207	0.0209	0.0608	0.7030	0.0260	0.0260	0.0500	0.5632	0.0263	0.0257	0.0592	0.6627
	ToT	0.0281	0.0278	0.0625	0.7185	0.0285	0.0286	<b>0.0805</b>	<b>0.8414</b>	0.0323	0.0316	0.0340	0.3770
	FAMA	0.0243	0.0227	0.0593	0.7470	0.0292	0.0281	0.0585	0.6868	0.0307	0.0319	<b>0.1013</b>	<b>1.2209</b>
Ours		0.0411	0.0406	0.0741	0.8186	<b>0.0436</b>	<b>0.0425</b>	0.0661	0.7075	<b>0.0422</b>	<b>0.0408</b>	0.0737	0.8103
30	GP	0.0149	0.0144	0.0541	0.6154	0.0262	0.0245	0.1422	1.5026	0.0220	0.0194	0.0815	0.7696
	DSO	-0.0009	0.0017	-0.0605	-0.6320	0.0067	0.0057	0.0037	0.0378	0.0250	0.0230	0.0919	0.9969
	AlphaGen	0.0204	0.0190	0.0417	0.4695	0.0240	0.0257	0.0538	0.6088	0.0415	0.0400	0.1149	1.2440
	AlphaForge	0.0348	0.0355	0.1162	1.0065	0.0283	0.0274	0.1004	1.2230	0.0317	0.0280	0.0732	0.9253
	CoT	0.0198	0.0188	0.1050	1.2943	0.0274	0.0264	<b>0.1631</b>	<b>1.6693</b>	0.0276	0.0250	0.1042	1.0422
	ToT	0.0265	0.0262	0.0975	1.1829	<b>0.0355</b>	<b>0.0353</b>	0.1330	1.5599	0.0337	0.0327	0.0584	0.7347
	FAMA	0.0341	0.0338	0.1173	1.3005	0.0302	0.0300	0.0596	0.6626	0.0331	0.0327	0.1012	1.0366
Ours		<b>0.0361</b>	<b>0.0359</b>	<b>0.1259</b>	<b>1.3210</b>	0.0337	0.0340	0.1235	1.3931	<b>0.0423</b>	<b>0.0424</b>	<b>0.1315</b>	<b>1.4307</b>

Table 11: The predictive performance of MLP model trained on alphas mined by different methods on the CSI300 stock pool.

$\Delta T$	Method	Alpha Num = 10				Alpha Num = 50				Alpha Num = 100			
		IC	RankIC	AR	IR	IC	RankIC	AR	IR	IC	RankIC	AR	IR
10	GP	0.0621	0.0581	0.0965	0.8339	0.0627	0.0587	0.0956	0.8933	0.0744	0.0646	0.1054	1.0063
	DSO	0.0362	0.0368	0.0926	0.8602	0.0492	0.0466	0.1091	1.0195	0.0609	0.0559	0.1137	1.0720
	AlphaGen	0.0528	0.0511	0.0770	0.7751	<b>0.0828</b>	<b>0.0720</b>	0.0403	0.4799	0.0793	0.0693	0.0816	0.9041
	AlphaForge	<b>0.0702</b>	<b>0.0628</b>	0.0768	0.7029	0.0664	0.0601	0.0881	0.7687	0.0728	0.0631	0.0512	0.4607
	CoT	0.0556	0.0498	0.0450	0.4726	0.0638	0.0574	0.0607	0.6113	0.0670	0.0597	0.0602	0.5818
	ToT	0.0619	0.0573	0.0993	0.9592	0.0599	0.0558	0.0954	0.9419	0.0654	0.0585	0.0909	0.8889
	FAMA	0.0631	0.0595	0.1053	0.9234	0.0643	0.0596	0.1134	1.0222	0.0647	0.0584	0.1186	1.0708
Ours		0.0661	0.0603	<b>0.1096</b>	<b>1.0919</b>	0.0748	0.0677	<b>0.1418</b>	<b>1.3699</b>	<b>0.0804</b>	<b>0.0729</b>	<b>0.1393</b>	<b>1.3577</b>
30	GP	<b>0.0575</b>	0.0522	0.0996	1.1359	0.0674	0.0637	0.0559	0.7088	0.0706	0.0672	0.0496	0.6230
	DSO	0.0362	0.0368	0.1181	1.1052	0.0420	0.0397	0.1096	1.2138	0.0609	0.0559	0.1226	1.1488
	AlphaGen	0.0511	0.0502	0.1087	1.2971	0.0727	0.0634	0.0991	<b>1.4196</b>	0.0685	0.0686	0.0520	0.7458
	AlphaForge	0.0558	<b>0.0532</b>	0.1095	0.9708	0.0705	0.0638	0.1081	0.9519	0.0705	0.0611	0.0796	0.8501
	CoT	0.0471	0.0438	0.1030	1.1473	0.0693	0.0629	0.1110	0.9757	0.0743	0.0661	0.1118	1.0872
	ToT	0.0475	0.0458	0.1340	1.4212	0.0643	0.0576	0.1097	1.1991	0.0758	0.0694	0.1147	1.1127
	FAMA	0.0543	0.0522	0.1136	0.9774	0.0550	0.0523	0.1340	1.2054	0.0557	0.0523	0.1226	1.0912
Ours		0.0529	0.0506	<b>0.1543</b>	<b>1.6588</b>	<b>0.0741</b>	<b>0.0673</b>	<b>0.1461</b>	1.4065	<b>0.0793</b>	<b>0.0723</b>	<b>0.1326</b>	<b>1.2598</b>

Table 12: The predictive performance of LightGBM model trained on alphas mined by different methods on the CSI1000 stock pool.

$\Delta T$	Method	Alpha Num = 10				Alpha Num = 50				Alpha Num = 100			
		IC	RankIC	AR	IR	IC	RankIC	AR	IR	IC	RankIC	AR	IR
10	GP	0.0578	0.0555	0.0922	0.8226	0.0606	0.0572	0.0807	0.6590	0.0659	0.0613	0.1031	0.9692
	DSO	0.0355	0.0349	0.1043	0.9314	0.0416	0.0405	0.0820	0.7957	0.0502	0.0469	0.0926	0.8907
	AlphaGen	0.0462	0.0463	0.0947	0.9076	0.0665	0.0619	0.0445	0.5072	0.0659	0.0597	0.0261	0.4004
	AlphaForge	<b>0.0694</b>	<b>0.0634</b>	0.1060	0.9916	0.0633	0.0580	0.1086	0.9420	0.0632	0.0606	0.0631	0.5837
	CoT	0.0518	0.0480	0.0513	0.5369	0.0585	0.0544	0.0721	0.7412	0.0648	0.0593	0.0706	0.6824
	ToT	0.0606	0.0578	0.1118	1.0644	0.0575	0.0558	0.0934	0.9376	0.0637	<b>0.0622</b>	0.0971	0.9258
	FAMA	0.0620	0.0591	0.1045	0.9361	0.0611	0.0582	0.1011	0.9539	0.0635	0.0572	0.1041	1.0323
Ours		0.0609	0.0581	<b>0.1286</b>	<b>1.3039</b>	<b>0.0681</b>	<b>0.0622</b>	<b>0.1168</b>	<b>1.1843</b>	<b>0.0662</b>	0.0618	<b>0.1204</b>	<b>1.2868</b>
30	GP	<b>0.0555</b>	<b>0.0517</b>	0.1590	1.4989	0.0653	0.0606	0.1105	1.2125	0.0655	0.0638	0.0977	1.0269
	DSO	0.0312	0.0318	0.1171	1.0981	0.0338	0.0326	0.1336	1.3199	0.0542	0.0522	0.1056	0.9626
	AlphaGen	0.0505	0.0505	0.1436	1.3123	0.0664	0.0611	0.0766	1.0838	0.0640	0.0622	0.0569	0.8894
	AlphaForge	0.0541	<b>0.0517</b>	0.1071	0.9857	0.0642	0.0606	0.1134	1.0138	0.0717	0.0650	0.0667	0.6850
	CoT	0.0465	0.0440	0.1176	1.2683	0.0662	0.0620	0.1345	1.3550	0.0698	0.0615	0.1174	1.4363
	ToT	0.0466	0.0447	0.1322	1.3771	0.0608	0.0556	0.1220	1.4510	0.0690	0.0633	0.1116	1.2468
	FAMA	0.0532	0.0511	0.1224	1.1036	0.0538	0.0519	0.1374	1.2472	0.0571	0.0530	0.1327	1.2086
Ours		0.0514	0.0491	<b>0.1799</b>	<b>1.8021</b>	<b>0.0673</b>	<b>0.0647</b>	<b>0.1710</b>	<b>1.6268</b>	<b>0.0738</b>	<b>0.0710</b>	<b>0.1696</b>	<b>1.5695</b>

Table 13: The predictive performance of MLP model trained on alphas mined by different methods on the CSI1000 stock pool.

### **Alpha Portrait Generation Prompt**

As shown in Figure 15, the Alpha Portrait Generation Prompt is used to generate an alpha portrait based on information such as available data fields and operators. The alpha portrait is subsequently used to generate the corresponding alpha formula.

### **Alpha Formula Generation Prompt**

As shown in Figure 16, the Alpha Formula Generation Prompt is used to generate the corresponding alpha formula based on the provided alpha portrait and other relevant information. A formatted alpha formula facilitates correctness verification and the computation of alpha values.

### **Alpha Overfitting Risk Assessment Prompt**

As shown in Figure 17, the Alpha Overfitting Risk Assessment Prompt is used to assess the overfitting risk of an alpha based on its formula and refinement history information. We provide evaluation criteria within the prompt to assist the LLM in conducting a critical assessment.

### **Alpha Refinement Prompt**

As shown in Figure 18, the Alpha Refinement Prompt is used to refine the original alpha formula based on refinement suggestions, thereby generating an improved formula.

---

**Task Description:**

You are a quantitative finance expert specializing in factor-based investing. Please design an alpha factor used in investment strategies according to the following requirements, and then provide the content of the alpha in the required format.

**Available Data Fields:**

The following data fields are available for use:

```
{available_fields}
```

**Available Operators:**

The following operators are available for use:

```
{available_operators}
```

**Alpha Requirements:**

1. The alpha value should be dimensionless (unitless).
2. The alpha should incorporate at least two distinct operations from the "Available Operators" list to ensure it has sufficient complexity. Avoid creating overly simplistic alphas.
3. All look-back windows and other numerical parameters used in the alpha calculation MUST be represented as named parameters in the pseudo-code. These parameter names MUST follow Python naming conventions (e.g., `lookback_period`, `volatility_window`, `smoothing_factor`).
4. The alpha should have NO MORE than 3 parameters in total.
5. The pseudo-code should represent the alpha calculation step-by-step, using only the "Available Operators" and clearly defined parameters. Each line in the pseudo-code should represent a single operation.
6. Use descriptive variable names in the pseudo-code that clearly indicate the data they represent.
7. When designing alpha expressions, try to avoid including the following sub-expressions:

```
{freq_subtrees}
```

**Formatting Requirements:**

The output must be in JSON format with three key-value pairs:

1. **"name"**: A short, descriptive name for the alpha (following Python variable naming style, e.g., `price.volatility_ratio`).
2. **"description"**: A concise explanation of the alpha's purpose or what it measures. Avoid overly technical language. Focus on the intuition behind the alpha.
3. **"pseudo\_code"**: A list of strings, where each string is a line of simplified pseudo-code representing a single operation in the alpha calculation. Each line should follow the format: `variable_name = op_name(input=[input1, input2, ...], param=[param1, param2, ...])`, where:
  - `variable_name` is the output variable of the operation.
  - `op_name` is the name of one of the "Available Operators".
  - `input1, input2, ...` are input variables (either from "Available Data Fields" or previously calculated variables, cannot be of a numeric type).
  - `param1, param2, ...` are parameter names defined in the alpha requirements.

The format example is as follows:

```
{
    "name": "volatility_adjusted_momentum",
    "description": ".....",
    "pseudo_code": [.....]
}
```

---

Figure 15: Alpha Portrait Generation Prompt

---

**Task Description:**

Please design a quantitative investment alpha expression according to the following requirements.

**Available Data Fields:**

- The following data fields are available for use: {available\_fields}

**Available Operators:**

- The following operators are available for use: {available\_operators}

**Alpha Requirements:**

- {alpha\_portrait\_prompt}

**Formatting Requirements:**

1. Provide the output in JSON format.
2. The JSON object should contain two fields: "formula", and "arguments".
  - "formula": Represents the mathematical expression for calculating the alpha.
  - "arguments": Represents the configurable parameters of the alpha.
3. "formula" is a list of dictionaries. Each dictionary represents a single operation and must contain four keys: "name", "param", "input", and "output".
  - "name": The operator's name (a string), which MUST be one of the operators provided in the "Available Operators" section.
  - "param": A list of strings, representing the parameter names for the operator. These parameter names MUST be used as keys in the "arguments" section.
  - "input": A list of strings, representing the input variable names for the operator. These MUST be data fields from the "Available Data Fields" or output variables from previous operations in the "formula", cannot be of a numeric type.
  - "output": A string, representing the output variable name for the operator. This output can be used as an input for subsequent operations.
4. "arguments" is a list of dictionaries. Each dictionary represents a set of parameter values for the alpha.
  - The keys of each dictionary in "arguments" MUST correspond exactly to the parameter names defined in the "param" lists of the "formula".
  - The values in each dictionary in "arguments" are the specific numerical values for the parameters.
5. You may include a maximum of 3 sets of parameters within the "arguments" field.
6. The parameter value that indicates the length of the lookback window (if applicable) must be within the {window\_range} range.
7. Ensure that the alpha expression is both reasonable and computationally feasible.
8. Parameter names should be descriptive and follow Python naming conventions (e.g., window\_size, lag\_period, smoothing\_factor). Avoid using single characters or numbers as parameter names.
9. Refer to the following example:

```
{  
    "formula": [.....],  
    "arguments": [.....]  
}
```

---

Figure 16: Alpha Formula Generation Prompt

---

**Task: Critical Alpha Overfitting Risk Assessment**

Critically evaluate the overfitting risk and generalization potential of the provided quantitative investment alpha, based on its expression and refinement history. Your assessment must focus on whether complexity and optimization appear justified or are likely signs of overfitting.

**Input:**

- **Alpha Expression:**

```
{alpha_formula}
```

- **Refinement History:**

```
{refinement_history}
```

**Evaluation Criteria:****1. Justified Rationale vs. Complexity:**

**Critique:** Is the complexity of the alpha expression plausibly justified by an inferred economic rationale, or does it seem arbitrary/excessive, suggesting fitting to noise?

**2. Principled Development vs. Data Dredging:**

**Critique:** Does the refinement history indicate hypothesis-driven improvements, or does it suggest excessive optimization and curve-fitting (e.g., frequent, unjustified parameter tweaks)?

**3. Transparency vs. Opacity:**

**Critique:** Is the alpha's logic reasonably interpretable despite its complexity, or is it opaque, potentially masking overfitting?

**Scoring & Output:**

- Assign a single **Overfitting Risk Score** from 0 to 10.
  - **10 = Very Low Risk** (High confidence in generalization)
  - **0 = Very High Risk** (Low confidence in generalization)
- Use the full 0-10 range to differentiate risk levels effectively.
- Provide a concise, one-sentence **Justification** explaining the score, citing the key factors from the criteria.
- Format the output as JSON, like the examples below:

**Example JSON Outputs:**

```
{  
    "reason": "Complexity is justified by a strong rationale; principled refinement  
    history suggests low risk.",  
    "score": 9  
}  
  
{  
    "reason": "Plausible rationale, but some expression opacity and parameter tuning in  
    history indicate moderate risk.",  
    "score": 5  
}  
  
{  
    "reason": "High risk inferred from opaque expression lacking clear rationale,  
    supported by history showing excessive tuning.",  
    "score": 1  
}
```

---

Figure 17: Alpha Overfitting Risk Assessment Prompt

---

**Task Description:**

There is an alpha factor used in quantitative investment to predict asset price trends. Please improve it according to the following suggestions and provide the improved alpha expression.

**Available Data Fields:**

The following data fields are available for use: {available\_fields}

**Available Operators:**

The following operators are available for use: {available\_operators}

**Alpha Suggestions:**

1. The alpha value should be dimensionless (unitless).
2. All look-back windows and other numerical parameters used in the alpha calculation MUST be represented as named parameters in the pseudo-code. These parameter names MUST follow Python naming conventions (e.g., lookback\_period, volatility\_window, smoothing\_factor).
3. The alpha should have NO MORE than 3 parameters in total.
4. The pseudo-code should represent the alpha calculation step-by-step, using only the "Available Operators" and clearly defined parameters. Each line in the pseudo-code should represent a single operation.
5. Use descriptive variable names in the pseudo-code that clearly indicate the data they represent.
6. When designing alpha expressions, try to avoid including the following sub-expressions: {freq\_subtrees}

**Original alpha expression:**

{origin\_alpha\_formula}

**Refinement suggestions:**

NOTE: The following improvement suggestions do not need to be all adopted; they just need to be considered and reasonable ones selected for adoption.

{refinementSuggestions}

**Formatting Requirements:**

The output must be in JSON format with three key-value pairs:

1. **"name"**: A short, descriptive name for the alpha (following Python variable naming style, e.g., price\_volatility\_ratio).
2. **"description"**: A concise explanation of the alpha's purpose or what it measures. Avoid overly technical language. Focus on the intuition behind the alpha.
3. **"pseudo\_code"**: A list of strings, where each string is a line of simplified pseudo-code representing a single operation in the alpha calculation. Each line should follow the format: variable\_name = op\_name(input=[input1, input2, ...], param=[param1, param2, ...]), where:
  - variable\_name is the output variable of the operation.
  - op\_name is the name of one of the "Available Operators".
  - input1, input2, ... are input variables (either from "Available Data Fields" or previously calculated variables, cannot be of a numeric type).
  - param1, param2, ... are parameter names defined in the alpha requirements.

The format example is as follows:

```
{  
    "name": "volatility_adjusted_momentum",  
    "description": ".....",  
    "pseudo_code": [.....]  
}
```

---

Figure 18: Alpha Refinement Prompt