Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Review:

A survey on large language model-based alpha mining

Junjie ZHANG[1], Shuoling LIU[2], Tongzhe ZHANG[2], Yuchen SHI[‡2,3]

[1]*College of Computing and Data Science, Nanyang Technological University, Singapore 639798, Singapore*

[2]*E Fund Management Co., Ltd., Guangzhou 510000, China*

[3]*Department of Industrial Systems Engineering and Management, National University of Singapore, Singapore 119077, Singapore*

E-mail: junjie.zhang@ntu.edu.sg; liushuoling@efunds.com.cn; zhangtongzhe@efunds.com.cn; shiyuchen@efunds.com.cn

Received June 7, 2025; Revision accepted Sept. 3, 2025; Crosschecked Sept. 29, 2025

Abstract: Alpha mining, which refers to the systematic discovery of data-driven signals predictive of future cross-sectional returns, is a central task in quantitative research. Recent progress in large language models (LLMs) has sparked interest in LLM-based alpha mining frameworks, which offer a promising middle ground between human-guided and fully automated alpha mining approaches and deliver both speed and semantic depth. This study presents a structured review of emerging LLM-based alpha mining systems from an agentic perspective, and analyzes the functional roles of LLMs, ranging from miners and evaluators to interactive assistants. Despite early progress, key challenges remain, including simplified performance evaluation, limited numerical understanding, lack of diversity and originality, weak exploration dynamics, temporal data leakage, and black-box risks and compliance challenges. Accordingly, we outline future directions, including improving reasoning alignment, expanding to new data modalities, rethinking evaluation protocols, and integrating LLMs into more general-purpose quantitative systems. Our analysis suggests that LLM is a scalable interface for amplifying both domain expertise and algorithmic rigor, as it amplifies domain expertise by transforming qualitative hypotheses into testable factors and enhances algorithmic rigor for rapid backtesting and semantic reasoning. The result is a complementary paradigm, where intuition, automation, and language-based reasoning converge to redefine the future of quantitative research.

Key words: Alpha mining; Quantitative investment; Large language models (LLMs); LLM agents; Fintech
https://doi.org/10.1631/FITEE.2500386 **CLC number:** TP391; TP18

1 Introduction

The construction of quantitative investment strategies is an inherently iterative and data-driven process, involving stages such as data preprocessing, alpha factor design, model training, portfolio optimization, trade execution, and performance attribution. At the core of this process lies alpha mining, the systematic discovery of signals that predict future cross-sectional returns. From a computational perspective, an alpha refers to any variable, either constructed or observed, which can be used to rank financial assets by their expected relative performance. Rather than forecasting exact price levels, an alpha aims to distinguish between likely outperformers and underperformers over a given horizon. A key feature of alpha mining is its practical orientation. Therefore, in this paper, we adopt a broad and practical definition: an alpha is any data-derived signal that exhibits empirical predictive power for excess returns. It may originate from financial ratios, technical indicators, sentiment scores, or other structured or unstructured information sources (Jegadeesh and

‡ Corresponding author
ORCID: Junjie ZHANG, https://orcid.org/0000-0001-7962-0680; Yuchen SHI, https://orcid.org/0000-0002-1885-8043

Titman, 1993; Chen HL et al., 2014). This view extends beyond the classical asset pricing notion, where the alpha denotes returns unexplained by known risk factors (Fama and French, 1993; Cochrane, 2011). In addition, the alpha mining research field concentrates mostly on single-factor alphas, i.e., structured, interpretable signals that encode specific return hypotheses. These alphas are typically defined in symbolic or rule-based form, and may depend on one or multiple features. Compared to black-box models that extract latent structure from high-dimensional inputs (Gu et al., 2020; Zhang Q et al., 2022), single-factor alphas provide transparency, control, and attribution. Each single-factor alpha can be tested independently, evaluated statistically, and combined modularly to form larger alpha libraries. These properties make them suitable for systematic validation and scalable deployment.

Alpha mining methodologies have evolved through multiple stages. Traditionally, the process has relied on human-driven alpha design, where candidate signals are manually constructed and peer-reviewed based on economic theory or domain expertise. Despite growing automation, such handcrafted alphas remain a cornerstone of quantitative research, valued for their interpretability, domain alignment, and theoretical grounding (Harvey et al., 2016; Kent et al., 2020). With the surge of artificial intelligence (AI), algorithm-based alpha mining emerges as a scalable alternative. Early systems leverage heuristic search (Mirjalili, 2019; Real et al., 2020), automated feature construction (Zhang TP et al., 2023), deep learning (Shi H et al., 2025), and reinforcement learning (Yu S et al., 2023) to uncover patterns across large feature spaces. These approaches expand the search frontier and reveal signals that often elude manual exploration. Empirical studies have shown that algorithmically mined alphas can match or exceed the out-of-sample performance of manually engineered ones (Chen AY et al., 2022), though often at the cost of transparency, interpretability, and operational robustness. More recently, advances in large language models (LLMs) and artificial general intelligence (AGI) have introduced a new paradigm. Frontier models such as generative pre-trained Transformer (GPT)-4 (OpenAI, 2023), Claude 3 (Anthropic, 2024), Gemini 1.5 (Gemini Team of Google, 2024), and DeepSeek R1 (DeepSeek-AI et al., 2025) exhibit strong capabilities in symbolic reasoning, code generation, and multi-stage instruction following. These shifts have led to the emergence of LLM-based alpha mining as a rapidly growing research area. The appeal of LLMs lies in three key advantages. First, LLMs can process and reason over unstructured information, such as news, earnings transcripts, and policy documents, beyond the reach of purely structured-data methods. They can also articulate the economic rationale behind a signal, producing natural language justifications useful for hypothesis refinement and regulatory review. Second, LLMs enable rapid generations of large numbers of candidate factors, accelerating the alpha discovery process relative to traditional human-led formulation. Third, LLMs support interactive, prompt-driven workflows that allow users to iteratively refine ideas in natural language, bridging the gap between domain intuition and executable implementation. Taken together, these capabilities position LLMs as a promising middle ground between human-guided design and fully automated search, offering both generative speed and semantic depth.

Although several surveys have examined the intersection of LLMs and financial AI, existing works have yet to systematically address the pivotal question: How can LLMs concretely benefit alpha mining? Notably, Guo J et al. (2024) proposed the paradigm of Quant 4.0, emphasizing automation, explainability, and knowledge-driven AI, but their work predates the widespread emergence of LLMs and does not address their role in alpha discovery. Similarly, Ding et al. (2024) offered a comprehensive review of LLM agents in financial trading, covering agent architectures, data modalities, and evaluation metrics. While they acknowledged alpha mining as an essential component of financial trading workflows, it was only briefly touched upon, with no dedicated synthesis. More broadly, Nie et al. (2024) and Cao BK et al. (2025) surveyed LLM applications in finance and quantitative investment, spanning sentiment analysis, forecasting, and reasoning, without zooming in on alpha mining as a distinct problem domain. To bridge this gap, this survey conducts a focused and in-depth review of LLM-based alpha mining, aiming to provide the first comprehensive synthesis in this emerging domain. Through targeted keyword searches, including LLM alpha mining, financial signal generation with LLMs, LLM quantitative trading, and LLM factor

discovery, we have identified and analyzed eight representative papers published between 2023 and 2025. We will analyze these recent academic efforts and assess whether LLMs represent a breakthrough in alpha mining research or a tool whose promise remains largely aspirational.

2 Formulations and engineering pathways of alpha mining systems

2.1 Definition and evaluation of alpha

Consider a financial market with n assets observed over T discrete time periods. Let $r_{i,t}$ denote the realized return of asset $i \in \{1, 2, \ldots, n\}$ at time $t \in \{1, 2, \ldots, T\}$. Let $\boldsymbol{x}_{i,t} \in \mathbb{R}^m$ be the feature vector associated with asset i at time t, consisting of m observable variables such as historical returns, fundamental indicators, technical signals, and macroeconomic data.

We define alpha as a function $\alpha : \mathbb{R}^m \to \mathbb{R}$ that maps features to a scalar prediction. The predicted return at time $t+1$ for asset i is

$$\hat{r}_{i,t+1} = \alpha(\boldsymbol{x}_{i,t}). \tag{1}$$

Therefore, the objective of alpha mining is to construct a function α such that $\hat{r}_{i,t+1}$ aligns with the realized return $r_{i,t+1}$. To evaluate an alpha, we consider three dimensions: effectiveness, stability, and practicality. Each corresponds to a distinct property of signal quality.

1. Effectiveness

This measures the predictive power of α in forecasting cross-sectional returns. A standard metric is the information coefficient information ratio (ICIR), defined as

$$\mathrm{ICIR} = \mathbb{E}[\mathrm{IC}_t]/\mathrm{Std}[\mathrm{IC}_t], \tag{2}$$

where $\mathbb{E}$ refers to expectation calculation, Std refers to standard deviation, and the information coefficient at time t, $\mathrm{IC}_t = \mathrm{Corr}(\alpha_t, r_{t+1})$, is the cross-sectional Pearson correlation between factor values and future returns. IC refers to the information coefficient. A higher ICIR indicates stronger and more consistent predictive performance.

Another widely used metric is the factor return, typically measured as the return spread between top and bottom quantiles:

$$\mathrm{FactorReturn}_t = r_t^{Q_5} - r_t^{Q_1}, \tag{3}$$

where $r_t^{Q_5}$ and $r_t^{Q_1}$ denote the average returns of the top and bottom quintile portfolios, respectively. Other metrics include RankIC (Spearman correlation) and hit ratio (accuracy of directional predictions).

2. Stability

This refers to the temporal and cross-sectional robustness of the alpha. The decay profile is a common diagnostic, which evaluates IC at multiple horizons h, and is used to assess the persistence of signal strength:

$$\mathrm{IC}_h = \mathrm{Corr}(\alpha_t, r_{t+h}). \tag{4}$$

Besides, orthogonality to known risk factors (e.g., market, size, and value) is often measured to ensure independence. Additional diagnostics include the standard deviation of IC, stability of factor returns, and robustness across subsamples and market regimes.

3. Practicality

This captures the implementability of the alpha in real-world settings. A widely used metric is the turnover ratio, which reflects trading costs due to position changes:

$$\mathrm{Turnover}_t = \frac{1}{2}\sum_{i=1}^{N} |w_{i,t} - w_{i,t-1}|, \tag{5}$$

where $w_{i,t}$ is the weight of asset i at time t, and N is the number of assets. Other considerations include factor breadth, the proportion of assets for which the factor is defined, and cost-adjusted returns that account for slippage, transaction costs, and market impact.

2.2 A general alpha mining pipeline

We outline a general alpha mining pipeline that applies across human-driven, algorithmic, and LLM-based workflows. As illustrated in Fig. 1, the process comprises four core modules: miner, implementor, backtester, and evaluator. These components are supported by a centralized data and knowledge base, which aggregates structured financial datasets (e.g., prices, fundamentals, and alternative sources) and domain knowledge (e.g., research reports, expert insights, and narrative priors). This base serves as a queryable substrate for informed alpha design. In addition, a dedicated human-system interaction layer facilitates collaboration across human experts, algorithmic tools, and LLMs throughout the pipeline.

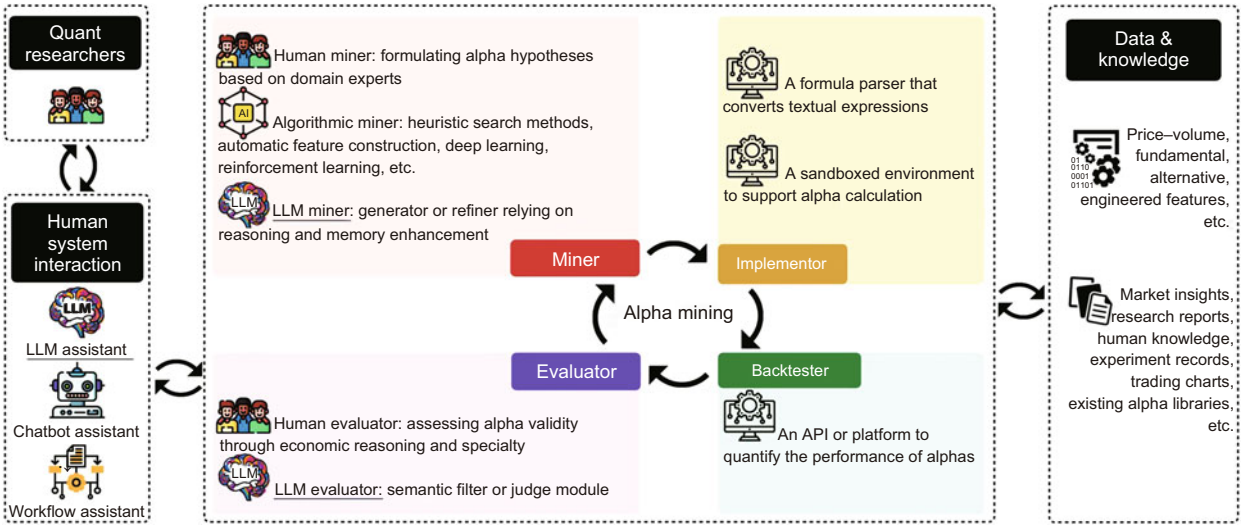

Fig. 1 A general alpha mining pipeline integrating human experts, algorithmic tools, and LLMs

We examine the roles and responsibilities of each core module in detail. These components form a loosely coupled pipeline, where each module performs a distinct function but remains interoperable with the others.

1. Miner

The miner generates and refines candidate alpha signals based on historical data and domain priors. It may operate via deductive reasoning or systematic exploration, e.g., perturbation, recombination, or constraint-based search. Outputs are typically expressed as symbolic formulas or executable code, supporting both abstraction and direct evaluation.

2. Implementor

The implementor transforms generated alpha logic into concrete, computable signals. It binds the required data inputs, handles preprocessing, and validates execution. For symbolic alphas, this involves expression parsing and feature resolution; for code-based alphas, it ensures secure and syntactically correct execution in sandboxed environments.

3. Backtester

The backtester evaluates historical performance for each alpha. It computes predictive and operational metrics (as defined in Section 2.1), and summarizes results under realistic market assumptions, forming the empirical foundation for downstream selection.

4. Evaluator

The evaluator integrates statistical outputs with economic reasoning. It assesses signal plausibility, robustness, and deployment feasibility. Operating autonomously or in human-in-the-loop mode, it scores candidate alphas and directs iterative refinement by prioritizing promising signals.

2.3 LLM-enhanced alpha mining pipeline

Having established the general architecture of an alpha mining pipeline, we turn our focus to the unique contributions of LLMs, which exhibit a versatile capability to operate across multiple stages of the pipeline. In the following, we first explain how current LLM-based alpha mining frameworks are organized, and then analyze how LLMs enhance each module in the alpha mining pipeline.

From the perspective of AGI, current LLM-based alpha mining frameworks coalesce around an agentic design that is essentially augmented by a triadic structure: tool, memory, and planning (Weng, 2023). While conceptually general, these elements are functionally asymmetric in the specific context of alpha mining. Planning is deliberately minimal. The alpha mining process typically follows a predetermined research–implementation–evaluation loop, leaving little room for autonomous decomposition or dynamic scheduling. In contrast, tool integration is indispensable. All frameworks couple LLMs with dedicated implementors and backtesters, which handle numerical evaluation beyond the model's generative scope. This coupling forms the backbone of the reasoning–feedback loop. Memory serves as the crucial bridge between one-shot generation

and sustained discovery. At its simplest, basic systems operate with prompt-only memory, relying entirely on internal model priors and static input, which limits adaptivity and generalization. More advanced systems incorporate retrieval-augmented memory from alpha libraries or structured knowledge bases, while some deploy persistent memory to regulate novelty, enforce constraints, and avoid redundancy across generations. Beyond the triadic infrastructure, the core differentiation among frameworks lies in how they shape the LLM's reasoning process. This involves shaping the model's generative process through structured workflows and hypothesis constraints. Simpler frameworks treat the LLM as a one-pass generator, producing candidate signals from static prompts. In contrast, ensemble workflows organize interaction across LLMs, humans, and auxiliary tools. The most structured designs deploy multi-agent systems, decomposing the alpha mining task into interpretable subroles, such as idea generation, formula synthesis, and semantic evaluation, each managed by specialized agents within the LLM framework. In summary, while research is still at its early exploratory stage, we observe a clear trend toward converging on a stable agentic template: an LLM scaffolded by tools, memory, and modular workflows. This convergence does not reflect stagnation but rather conceptual consolidation. As seen in the evolution of retrieval-augmented generation (RAG) systems, convergence often precedes scale and refinement. The key innovation lies not in architectural novelty, but in how each framework disciplines the model's generative capacity with domain-aware constraints, thus transforming a general-purpose LLM into a useful alpha mining framework.

Having clarified the agentic design of LLM-based alpha mining frameworks, we show Fig. 1 to illustrate how the general alpha mining pipeline is enhanced by LLMs. Essentially, LLMs enhance the alpha mining pipeline by flexibly occupying three principal roles—miner, evaluator, and interactive assistant. As miners, LLMs expand the hypothesis space by synthesizing novel alpha expressions from domain knowledge, retrieved signals, or natural language prompts and further refine the hypothesis via memory-augmented reasoning, thereby augmenting both the diversity and creativity of candidate factors. As evaluators, LLMs serve as semantic filters that assess the plausibility, originality, and interpretability of generated signals, offering a complementary perspective to purely statistical validation. Finally, LLMs can serve directly as interactive assistants. As can be seen in Fig. 1, human–system interaction falls into three paradigms. Workflow assistants provide stability but are rigid and brittle in open-ended tasks. ChatBot assistants offer structured interaction, but often lack depth and adaptability. Among them, direct LLM assistants stand out as they enable fluid dialogue, fast iteration, and context-aware reasoning. By embedding these multi-role capabilities into each module, LLMs not only improve the efficiency of the alpha mining process but also make it more interpretable, scalable, and adaptive to real-world constraints.

3 Dissecting the functional roles of LLMs in alpha mining

Based on the previous descriptions, in this section, we explain more explicitly how LLMs can be cast as the above-mentioned functional components by reviewing the representative works. These roles are not mutually exclusive. Most recent frameworks adopt hybrid configurations that integrate multiple roles in a coordinated workflow. In essence, the role assignment reflects a trade-off between automation and control. We summarize the reviewed works in Table 1.

3.1 LLMs as the miner

When acting as a miner, the LLM maps raw knowledge and structured data into candidate alpha functions:

$$\mathrm{LLM}_{\mathrm{Miner}} : (\mathcal{K}, \mathcal{D}) \to \alpha, \tag{6}$$

where $\mathcal{K}$ represents domain knowledge and $\mathcal{D}$ denotes historical datasets. The output α may take the form of a symbolic expression or code snippet that is suitable for further evaluation. Typical instantiations include formula generation, hypothesis translation, and feature recombination. This role is foundational in nearly all reviewed systems. Most of the memory and cognitive workflows reviewed are designed to enhance LLM's capability as a miner. The simplest systems, such as those in Cheng and Tang (2024) and Wang YN et al. (2024), treat the LLM as a one-pass generator. They use static prompts and

Table 1 Summary of representative LLM-based alpha mining frameworks

Framework	Open source	LLM used	Data type	Knowledge base	LLM role
Alpha-GPT (Wang SZ et al., 2023)	No	GPT-3.5-turbo-16k	Price–volume	Existing alpha library and historical experiment records	Miner and interactive assistant
Cheng-Tang-2023 (Cheng and Tang, 2024)	No	ChatGPT	Price–volume	None	Miner and evaluator
GPT-signal (Wang YN et al., 2024)	Planned	GPT-4	Fundamental	None	Miner
Kou-2024 (Kou et al., 2024)	Yes	ChatGPT	Price–volume	Multimodal data: market dynamics, financial reports, trading charts, etc.	Miner
QuantAgent (Wang SZ et al., 2024)	No	GPT-4	Price–volume	Historical experiment records	Miner and evaluator
FAMA (Li ZW et al., 2024)	No	GPT-3.5 (text-davinci-002)	Price–volume	Existing alpha library	Miner
R&D-Agent (Li YT et al., 2025)	Yes	GPT-4o-(mini), o3-(mini), GPT-4.1, GPT-4-turbo	Price–volume	Historical experiment records	Miner, evaluator, and interactive assistant
AlphaAgent (Tang et al., 2025)	No	GPT-3.5-turbo	Price–volume	Human knowledge, research reports, and market insights	Miner and evaluator
LLM-Powered MCTS (Shi Y et al., 2025)	No	GPT-4.1	Price–volume	Existing alpha library and historical experiment records	Miner and evaluator
Chain-of-Alpha (Cao L et al., 2025)	No	GPT-4o, DeepSeek-V3, Qwen3-32B	Price–volume	Existing alpha library and historical experiment records	Miner

MCTS: Monte Carlo tree search; FAMA: foundation for the advancement of monetary affairs

rely on internal model priors to produce alpha expressions directly from textual instructions. These approaches are lightweight and intuitive, but often struggle with adaptivity and generalization. More structured approaches emphasize memory and reasoning, adopting a more mature agentic design. For instance, Alpha-GPT (Wang SZ et al., 2023) includes user feedback loops and auxiliary algorithmic tools that help exploit and refine promising ideas. The system employs a prompt to interpret natural language inputs from quantitative researchers, and translates them into executable and high-quality alpha expressions by integrating knowledge extracted from the existing alpha library. The inclusion of user feedback loops and auxiliary algorithmic tools further refines the alpha mining strategies. R&D-Agent (Chen HT et al., 2024; Yang et al., 2024) focuses on data-centric automatic research and development (R&D) by benchmarking operations that extract methods from raw information, such as financial reports and papers, facilitating the implementation of methods through code. It further tracks research history across development stages, enabling conversational editing and hypothesis iteration. FAMA (Li ZW et al., 2024) is a neural-symbolic framework that leverages external knowledge through two key components: cross-sample selection (CSS) and chain-of-experience (CoE). CSS mitigates homogeneity in factor generation by incorporating diverse, low-correlation factors as contextual samples, while CoE guides the LLM to explore novel factor paradigms by using past successful mining paths as experiential prompts. QuantAgent (Wang SZ et al., 2024) implements a writer—judge loop, where the mining phase is constrained by feedback from an internal judge that scores for novelty, performance alignment, and economic rationale. Kou et al. (2024) extracted alpha factors from multimodal financial data. Seed alphas are assigned with confidence scores based on predictive quality and market status. A dynamic weight-gating mechanism then selects and assigns weights, enabling the creation of an adaptive and context-aware composite alpha formula. AlphaAgent (Tang et al., 2025) introduces a multi-agent framework with three key mechanisms: (1) originality enforcement through similarity measures against existing alpha libraries, (2) hypothesis–factor alignment via LLM-evaluated semantic consistency

between market hypotheses and generated factors, and (3) complexity control to prevent overfitting. These mechanisms collectively guide the alpha generation process to balance originality, financial rationale, and adaptability to evolving market conditions, mitigating the risk of alpha decay.

3.2 LLMs as the evaluator

When acting as an evaluator, the LLM interprets both the statistical performance and the semantic structure of a candidate alpha to provide diagnostic feedback or filtering. We formalize this role as

$$\mathrm{LLM}_{\mathrm{Eval}} : \{\mathcal{M}_k(\alpha), \mathcal{T}(\alpha)\}_{k=1}^{K} \to \mathcal{S}(\alpha), \quad (7)$$

where $\mathcal{M}_k(\alpha)$ denotes traditional evaluation metrics such as IC, Sharpe ratio, or turnover, and $\mathcal{T}(\alpha)$ represents natural language descriptions or code structure capturing the semantic logic of α. The resulting $\mathcal{S}(\alpha)$ reflects a synthesized assessment based on quantitative indicators and qualitative reasoning. This formulation supports evaluator behaviors such as performance-guided selection, semantic alignment checking, and explanation generation. Rather than relying solely on statistical metrics, systems increasingly invoke the LLM to assess the plausibility, novelty, or coherence of a factor. For example, in QuantAgent, a judge module evaluates signals based on expected performance and prior knowledge (Wang SZ et al., 2024). In AlphaAgent, evaluation is framed as alignment checking, verifying that the factor remains faithful to the underlying economic rationale (Tang et al., 2025). More generally, LLMs are often used to generate natural language explanations that support human decision-making (Wang SZ et al., 2023).

3.3 LLMs as the interactive assistant

Beyond fixed tasks, LLMs can act as a natural language interface between the system and the researcher. Beyond generation and evaluation, it supports interpretation, guided prompting, and task-level feedback:

$$\mathrm{LLM}_{\mathrm{Chat}} : (\mathrm{Prompt}_{\mathrm{user}}, \mathcal{C}) \to \mathrm{Action}_{\mathrm{next}}, \quad (8)$$

where $\mathrm{Prompt}_{\mathrm{user}}$ denotes the prompt inserted by the user, $\mathcal{C}$ denotes context (e.g., current alpha, performance logs, and external documents), and $\mathrm{Action}_{\mathrm{next}}$ may include modifying formulas, querying knowledge, or planning subsequent evaluations. This aligns with agentic workflows such as human-in-the-loop alpha tuning or co-design. Together, these roles frame the LLM to support both autonomous alpha discovery and human-guided iteration, enabling a hybrid intelligence paradigm that bridges machine reasoning and domain expertise. For example, Alpha-GPT exemplifies this mode, translating user intuitions into factor code while offering justifications (Wang SZ et al., 2023). R&D-Agent builds on this further by logging task iterations and enabling conversational refinement (Chen HT et al., 2024; Yang et al., 2024).

4 Reported LLM-based alpha mining results

After presenting the alpha mining pipeline and the functional roles of LLMs, one open question remains: How good are the alphas mined by LLMs in practice? To address this, we compile the reported experiment results from the literature, summarized in Table 2. All the reported backtesting results follow the academic convention of daily-frequency evaluation, except that LLM-Powered MCTS calculates 10-d and 30-d forward return and Chain-of-Alpha calculates 10-d forward return. Note that the literature often adopts very different choices of assets, backtest periods, evaluation metrics, and baselines in empirical experiments and case studies. As a result, while we list the reported results as clearly as possible for completeness, the comparability across studies is inherently limited.

Across the reported works, we can observe several consistent patterns. First, model choice matters: most experiments rely on the GPT series as the base model. GPT-4 based approaches generally report higher Sharpe ratios and annualized returns than GPT-3.5 based ones. Only Cao BK et al. (2025) disclosed results with more recent and mature models (GPT-4o, DeepSeek-V3, and Qwen3-32B), and their comparison shows that these three models perform on par with each other. Second, advanced search or control mechanisms—such as dual-loop self-improvement (Wang SZ et al., 2024), MCTS guidance (Shi Y et al., 2025), and CoT style parallel exploration (Cao BK et al., 2025)—tend to yield stronger IC/RankIC stability and higher

Table 2 LLM-based factor mining experiments and baseline comparisons

Framework	Data scope	Disclosed backtest metric	Comparison with baseline
Alpha-GPT (Wang SZ et al., 2023)	US S&P500 (2012–2021)	Single factor IC=0.020–0.025; annualized return=5%–10%	IC doubled compared with LLM-only generation (0.010→0.020); AR improved from marginally positive to 5%–10%
Cheng-Tang-2023 (Cheng and Tang, 2024)	CRSP US Stock (2021–2022)	Single factor AR=66.16%, Sharpe=4.49; equal-weighted GPT factor portfolio AR=88%, Sharpe=2.46	Excess returns unexplained by Fama–French five-factor model; superior to heuristic factor design in novelty and efficiency
GPT-signal (Wang YN et al., 2024)	US S&P500 sectoral data (healthcare, IT, energy) (2016–2020)	IC improved by 3%–5%; annualized return improved by 5%–8%; portfolio Sharpe from 1.2 to >1.5; strongest gains in IT and healthcare; development time reduced by 30%	IC and Sharpe improved by 30% compared with traditional financial-signal models
QuantAgent (Wang SZ et al., 2024)	China A-share, 500 stocks (2023)	IC from 0.009 to ∼0.018 after five self-improvement cycles; Sharpe increased from 0.85 to >1.25; AR improved by ∼5%–7% compared to a static baseline	Dual-loop (inner+outer) mechanism outperforming ablated single-loop settings: without the outer loop IC stagnated at ∼0.012 and without the inner loop Sharpe dropped to <1; overall performance nearly doubled relative to static factors
FAMA (Li ZW et al., 2024)	US S&P500 (2015–2022)	RankIC=0.054, RankICIR=0.485, annualized return ≈38.4%, Sharpe ≈6.67	Dual-loop (inner+outer) mechanism outperforming single-loop variants; IC stagnation (∼0.012) without outer loop; Sharpe <1 without inner loop; performance nearly doubled vs. static factors
Kou-2024 (Kou et al., 2024)	China SSE 50 (2021–2023)	cum. return 53.17%, Sharpe >1.5, MaxDD −7%	AR improved by +30%–40%; Sharpe improved by 0.5 compared to classical and deep (ALSTM, Transformer) models; superior to AlphaEvolve and RL-based alpha factories in return and drawdown control
AlphaAgent (Tang et al., 2025)	China CSI500, US S&P500 (2015–2024)	CSI500: AR=11.00%, IR=1.488, IC=0.0212, MaxDD=−9.36%; S&P500: AR=8.74%, IR=1.055, IC=0.0056, MaxDD=−9.10%	Superior AR, IR, and IC vs. LSTM, Transformer, LightGBM, TRA, StockMixer, AlphaForge, R&D-Agent, and LLM baselines (OpenAI-o1, DeepSeek-R1)
R&D-Agent-Quant (Li YT et al., 2025)	China CSI300 (2016–2024)	RD-Factor: IC=0.0497, IR=1.36, ARR=11.84%, MaxDD=−9.10%; RD-Model: IC=0.0469, IR=1.70, ARR=10.09%, MaxDD=−6.94%; RD-Agent (Q): IC=0.0532, IR=1.74, ARR=14.21%, MaxDD=−7.42%	Superior to Alpha101 factor library, GP-based methods, and earlier RD-Agent; also superior to deep models (LSTM/Transformer) on CSI300 in both IC and risk-adjusted return
LLM-Powered MCTS (Shi Y et al., 2025)	China CSI300, CSI1000 (2011–2024); US S&P500 (2015–2024)	CSI300: AR=8.20%, IR=0.94, IC=0.0420, RankIC=0.0395; CSI1000: AR=13.90%, IR=1.36, IC=0.0800, RankIC=0.0730; S&P500: IC=0.0132, RankIC=0.0130	Superior IC/RankIC/AR/IR vs. GP, DSO, AlphaGen, AlphaForge, CoT/ToT, FAMA, and AlphaAgent; consistent advantage on CSI300/CSI1000 and S&P500
Chain-of-Alpha (Cao L et al., 2025)	China CSI500 & CSI1000 (2010–2025)	CSI500: IC=0.0485, RankIC=0.0771, ICIR=0.3047, RankICIR=0.5013, AR=0.1324, IR=1.4178; CSI1000: IC=0.0672, RankIC=0.0902, ICIR=0.4630, RankICIR=0.6228, AR=0.1471, IR=1.4043	Superior to Alpha101/158/360, GP, DSO, AlphaGen, AlphaForge, and LLM baselines (LLM+CoT/ToT/MCTS)

ALSTM: attention-based long short-term memory; AR: autoregressive; CoT: chain-of-thought; CRSP: center for research in security prices; CSI: China securities index; DSO: direct search optimization; GP: genetic programming; IR: information ratio; LSTM: long short-term memory; SSE: Shanghai stock exchange; ToT: tree-of-thought; TRA: trading return analysis; cum.: cumulative

IR than single-pass prompting or static generation. Third, compared with traditional factor libraries (Alpha101/158/360), symbolic methods (GP and DSO), and machine learning baselines (LSTM, Transformer, and LightGBM), LLM-based methods almost universally show gains in commonly seen backtest metrics such as IC and IR, and generally produce more interpretable signals. Literature such

as Cheng and Tang (2024) points out that the excess returns of GPT-generated factors cannot be explained by the Fama–French five-factor model, indicating that these factors contain new information beyond traditional asset pricing frameworks.

At the same time, limitations remain. Reported results differ widely in data scope (US vs. China markets), evaluation frequency (daily vs. multi-day horizons), and baseline selection, making direct comparability limited. More importantly, some studies emphasize the analysis of single factors, while others focus on portfolio-level returns; some highlight comparisons with existing methods, whereas other experimental setups are designed to validate the advantages of their proposed LLM-based approaches rather than to perform a direct comparative benchmark against other specific alpha mining algorithms from the literature. In addition, current evidence remains constrained by model choice, as most studies rely on the GPT series with only limited exploration of alternatives such as DeepSeek or Qwen. Moreover, there has been no systematic discussion of how configurations, such as distillation, quantization, and extended context length, affect mining effectiveness. We highlight this lack of discussion as a research gap for future benchmark studies.

In summary, existing experiments demonstrate that LLMs already generate alpha signals that outperform traditional factor mining methods across multiple markets and benchmarks. However, the heterogeneity in experimental settings and the absence of standardized evaluation protocols mean that the current evidence is promising but still fragmented. Establishing unified benchmarks will be an important step for assessing the true effectiveness of LLM-based alpha mining.

5 Challenges

Despite the promising potential of LLMs in assisting alpha mining, several critical challenges remain, hindering their practical adoption in rigorous quantitative investment workflows.

1. Simplified performance evaluation

Most existing studies emphasize feasibility over rigor. Performance is often demonstrated using simplified settings, loose validation criteria, or selective backtests. Few works benchmark against established academic or industry-standard factors, nor do they quantify incremental value beyond naive baselines. In some cases, this is due to academic abstraction or confidentiality constraints; in others, it reveals limitations in handling the full rigor of production-grade alpha validation. As a result, the connection between research output and real-world utility remains weak.

2. Limited numerical understanding

LLMs' ability to handle complex numerical logic remains limited. While factor generation from price–volume data is common, extending to fundamentals or alternative datasets requires precise accounting logic, time alignment, and cross-sectional consistency—requirements that are often under-specified or ignored. More critically, fully end-to-end pipelines risk introducing information leakage, system-level coupling, and label contamination. These issues compromise factor stability and generalization, leading to inflated backtest results and poor live deployment outcomes.

3. Lack of diversity and originality

Many LLM-based frameworks have yet to demonstrate the ability to consistently generate diverse and novel alpha signals. Most empirical results report only a handful of candidate factors, without showing whether the system can support continued idea generation over time. A central challenge lies in how to integrate heterogeneous knowledge sources (such as academic finance literature, domain priors, and latent signals in unstructured data) and systematically transform them into valid, interpretable alpha expressions. While RAG offers a possible route, the finance domain poses unique barriers: Publicly available research tends to include qualitative ideas but lacks reproducible alpha construction details, and the natural language descriptions of trading strategies are often imprecise or ambiguous. Bridging the gap between vague narrative intent and executable factor logic remains non-trivial.

4. Weak exploitation dynamics

While LLMs are effective in creative ideation, they often fall short in iterative refinement. Empirically, many systems fail to exhibit consistent improvement across generations. Without explicit optimization objectives or structured feedback loops, the model's search process lacks direction. This reflects a deeper tension: Exploration is well supported by generative priors, but exploitation—extracting durable value—requires more disciplined control and training strategies.

5. Temporal data leakage

The risk of information leakage is particularly acute in finance. LLMs may have been pretrained on historical data, but robust factor validation typically spans a 10-year horizon or more. It becomes difficult to determine whether strong results reflect true signal discovery or inadvertent access to future-sensitive patterns seen during training. Without explicit decontamination mechanisms, this undermines claims of generalization and may lead to spurious conclusions.

6. Black-box risks and compliance challenges

Integrating black-box LLM inferences into quantitative workflows raises concerns over auditability, explainability, and regulatory compliance. Note that compared with conventional deep learning methods, LLMs already offer a partial remedy: Their ability to generate natural language explanations for discovered factors provides a novel channel for interpretability and human oversight. Therefore, the challenge is to bridge the gap between free-form explanations and rigorous, regulator-ready documentation. Without provenance tracking, structured diagnostics, and human-in-the-loop validation, deploying such signals in production remains problematic. Enhancing transparency and compliance will be essential before LLMs can be safely adopted in institutional investment pipelines.

6 Future directions

Looking ahead, several research directions can be pursued to address the current limitations and expand the frontier of LLM-driven alpha mining. Some of these directions respond directly to the challenges discussed earlier, while others reflect natural progressions as the field evolves beyond its nascent stage.

1. Enhancing reasoning structure through domain alignment

Based on the limitations discussed above, future work with more structured task guidance will appear to further improve LLM reasoning. In parallel, domain-specific fine-tuned models may be trained to improve LLMs' ability to interpret backtesting outputs, perform quantitative diagnostics, and reason over financially grounded metrics. Progress in this direction hinges on the construction of high-quality, labeled financial corpora and carefully designed reward functions that reflect investment-relevant metrics. A particularly promising direction lies in enhancing the numerical reasoning capabilities of LLMs, which are critical for expressing, manipulating, and validating financial signals in alpha mining. Recent studies have proposed several promising approaches. For instance, Srivastava et al. (2024) evaluated LLMs' mathematical reasoning on financial tabular datasets, highlighting the importance of tailored prompting techniques to improve performance in complex numerical tasks. Furthermore, Su et al. (2024) proposed NumLLM, a numeric-sensitive LLM fine-tuned on financial corpora, demonstrating enhanced understanding of financial texts involving numerical variables. These methodologies collectively suggest that integrating structured numerical reasoning and domain-specific adaptations into LLMs can significantly advance their effectiveness in alpha factor generation. A more comprehensive review on how foundation models and the fine-tuning process can empower financial applications can be seen in Chen LY et al. (2025).

2. Expanding into fundamental and alternative data domains

Extending current LLM-based frameworks to structured fundamental data (e.g., income statements and balance sheets) and alternative datasets (e.g., environmental, social, and governance (ESG) scores, satellite data, and transaction records) presents a rich avenue for future exploration, enabling a broader and more diversified alpha hypothesis space. The challenge lies in translating these heterogeneous inputs into formats that LLMs can reason over while preserving context, causality, and economic meaning. Recent studies have proposed several promising strategies for integrating structured and semi-structured data into LLM workflows. Wu et al. (2023) introduced BloombergGPT, a 50-billion-parameter language model trained on a diverse corpus of financial data, demonstrating superior performance in financial natural language processing (NLP) tasks. Mehra et al. (2022) developed ESGBERT, a domain-specific language model fine-tuned on ESG-related texts, achieving improved accuracy in ESG classification tasks. Additionally, Xia et al. (2024) presented a pipeline leveraging pre-trained language models for accurate ESG prediction, highlighting the effectiveness of domain-specific adaptations. These approaches suggest that augmenting LLMs with structured

data parsers, schema-aware embeddings, or external memory modules can enhance their capability to process non-textual financial inputs effectively. Moving toward multimodal signals will inevitably require more advanced fusion strategies (e.g., cross-modal embeddings and data fusion pipelines) and scalable training infrastructure, and we highlight this as an important but currently unexplored research gap for future work.

3. Revisiting backtesting and evaluation methodologies

The entry of LLMs offers an opportunity to rethink classical evaluation pipelines. Future frameworks may leverage LLMs to automate hypothesis formation, construct regime-aware benchmarks, or simulate counterfactual market conditions. These capabilities could modernize traditional backtesting pipelines and increase their flexibility under varying market regimes. Note that such contributions are complementary to, rather than substitutes for, the broader challenges of standardized validation, benchmarking, and industry comparability discussed earlier. Recent studies have proposed innovative approaches to enhance backtesting and evaluation methodologies using LLMs. For example, Tang et al. (2025) integrated LLM agents with regularization mechanisms to mitigate alpha decay, employing originality enforcement, hypothesis alignment, and complexity control. While this work does not directly establish new evaluation standards, it illustrates how LLM integration can expand the scope of evaluation practices. Building on this, future work should aim to formalize transparent and reproducible validation schemes aligned with established benchmarks, and to further explore how LLMs can enable regime-aware, context-sensitive, and more adaptive forms of backtesting. With such efforts, LLMs may help establish a new paradigm for hypothesis testing and signal validation in modern alpha mining systems.

4. Developing a comprehensive general-purpose quantitative investment framework

As LLMs evolve from narrow tools to generalized agents, a broader ambition emerges: the construction of comprehensive quantitative platforms that integrate macroeconomic signals, sentiment data, and multi-modal inputs—including tabular, textual, visual, and even audio information. Recent efforts have begun to explore these comprehensive platforms. Yuan et al. (2024) proposed Alpha-GPT 2.0, a quantitative investment framework that further encompasses crucial modeling and analysis phases in quantitative investment. Their framework demonstrates how LLMs can participate at multiple stages of the investment workflow—from interpreting qualitative narratives to generating interpretable alpha signals. Similarly, Papasotiriou et al. (2024) presented a large-scale evaluation of LLMs applied to equity stock rating tasks, highlighting their ability to blend structured fundamentals with financial language modeling. Their findings suggest that LLMs can support semi-systematic investment decisions that historically require deep analyst expertise. Looking further, Yu YY et al. (2024) introduced FinCon, a multi-agent LLM system enhanced with verbal reinforcement learning, enabling agents to reason collaboratively over complex financial scenarios. Their architecture supports interactive decision-making across modalities and agents, hinting at future intelligent platforms where LLMs orchestrate data ingestion, reasoning, and execution within unified pipelines. Together, these works point to an emerging vision: general-purpose quantitative systems where LLMs not only automate subtasks but also mediate among diverse informational sources, human decision-makers, and machine-learning-based execution engines. Realizing this vision will require advances in multi-modal learning, agent coordination, and domain alignment protocols.

7 Conclusions

This paper presents the first comprehensive survey of LLMs in the domain of alpha mining—an emerging intersection of natural language reasoning and quantitative finance. We begin by formalizing the notion of alpha as a structured, interpretable signal for cross-sectional return prediction, and outline a modular pipeline encompassing miner, implementor, backtester, and evaluator components. We then dissect how recent LLM-based systems—such as Alpha-GPT, QuantAgent, and AlphaAgent—adopt agentic designs to enhance different modules in the pipeline, with LLMs playing functional roles as miners, evaluators, and interactive assistants.

Despite promising capabilities, current frameworks face several challenges: Performance evaluation remains simplified, numerical reasoning is limited, original signal generation lacks depth,

exploitation dynamics are weak, and risks of temporal leakage persist. These limitations hinder LLMs from being deployed in production-grade workflows. To address these issues, we highlight four promising research directions: (1) enhancing LLM reasoning structures through domain alignment and fine-tuning, (2) expanding into fundamental and alternative data domains, (3) revisiting backtesting and evaluation methodologies, and (4) developing a comprehensive general-purpose quantitative investment framework that unifies human, algorithmic, and LLM capabilities.

As a broader discussion, we notice that the trajectory of LLM-based alpha mining mirrors the development trajectory of AGI: from emergent capability to structured reliability. The field stands at a promising threshold, with early systems proving feasible, and future ones poised to reshape the practice of alpha discovery at scale.

Ultimately, we argue that the value of LLMs in alpha mining lies not in replacing human or algorithmic intelligence but in amplifying it. Human expertise remains the most differentiated source of alpha, while LLMs offer a powerful interface for accelerating hypothesis validation. In tandem, algorithmic mining provides scale and consistency, whereas LLMs excel at cold-start exploration and semantic reasoning. This hybrid paradigm—merging domain intuition, computational rigor, and generative flexibility—represents a promising frontier for the next generation of quantitative research.

Contributors

Junjie ZHANG, Shuoling LIU, and Yuchen SHI designed the outline of the review. Junjie ZHANG and Yuchen SHI surveyed the literature. Junjie ZHANG, Yuchen SHI, and Tongzhe ZHANG drafted the paper. Yuchen SHI revised and finalized the paper.

Conflict of interest

Shuoling LIU is a guest editor of the Special Feature on Theories and Applications of Financial Large Models of *Frontiers of Information Technology & Electronic Engineering*; he was not involved with the peer review process of this paper. All the authors declare that they have no conflict of interest.

References

Anthropic, 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. Anthropic Research Report. https://assets.anthropic.com/m/61e7d27f8c8f5919/original/Claude-3-Model-Card.pdf [Accessed on June 21, 2025].

Cao BK, Wang SZ, Lin XY, et al., 2025. From deep learning to LLMs: a survey of AI in quantitative investment. https://doi.org/10.48550/arXiv.2503.21422

Cao L, Xi ZK, Liao L, et al., 2025. Chain-of-Alpha: unleashing the power of large language models for alpha mining in quantitative trading. https://doi.org/10.48550/arXiv.2508.06312

Chen AY, Lopez-Lira A, Zimmermann T, 2022. Does peer-reviewed research help predict stock returns? https://doi.org/10.48550/arXiv.2212.10317

Chen HL, De P, Hu Y, et al., 2014. Wisdom of crowds: the value of stock opinions transmitted through social media. *Rev Fin Stud*, 27(5):1367-1403. https://doi.org/10.1093/rfs/hhu001

Chen HT, Shen XJ, Ye ZQ, et al., 2024. RD2Bench: toward data-centric automatic R&D. Proc 13th Int Conf on Learning Representations, p.1-22.

Chen LY, Liu SL, Yan JP, et al., 2025. Advancing financial engineering with foundation models: progress, applications, and challenges. https://doi.org/10.48550/arXiv.2507.18577

Cheng YH, Tang K, 2024. GPT's idea of stock factors. *Quant Fin*, 24(9):1301-1326. https://doi.org/10.1080/14697688.2024.2318220

Cochrane JH, 2011. Presidential address: discount rates. *J Fin*, 66(4):1047-1108. https://doi.org/10.1111/j.1540-6261.2011.01671.x

DeepSeek-AI, Guo DY, Yang DJ, et al., 2025. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. https://doi.org/10.48550/arXiv.2501.12948

Ding H, Li YH, Wang JH, et al., 2024. Large language model agent in financial trading: a survey. https://doi.org/10.48550/arXiv.2408.06361

Fama EF, French KR, 1993. Common risk factors in the returns on stocks and bonds. *J Fin Econ*, 33(1):3-56. https://doi.org/10.1016/0304-405X(93)90023-5

Gemini Team of Google, 2024. Gemini 1.5: unlocking multimodal understanding across millions of tokens of context. https://doi.org/10.48550/arXiv.2403.05530

Gu SH, Kelly B, Xiu DC, 2020. Empirical asset pricing via machine learning. *Rev Fin Stud*, 33(5):2223-2273. https://doi.org/10.1093/rfs/hhz009

Guo J, Wang SZ, Ni LM, et al., 2024. Quant 4.0: engineering quantitative investment with automated, explainable, and knowledge-driven artificial intelligence. *Front Inform Technol Electron Eng*, 25(11):1421-1445. https://doi.org/10.1631/FITEE.2300720

Harvey CR, Liu Y, Zhu HQ, 2016. ... and the cross-section of expected returns. *Rev Fin Stud*, 29(1):5-68. https://doi.org/10.1093/rfs/hhv059

Jegadeesh N, Titman S, 1993. Returns to buying winners and selling losers: implications for stock market efficiency. *J Fin*, 48(1):65-91. https://doi.org/10.1111/j.1540-6261.1993.tb04702.x

Kent D, Lira M, Simon R, et al., 2020. The cross-section of risk and returns. *Rev Fin Stud*, 33(5):1927-1979. https://doi.org/10.1093/rfs/hhaa021

Kou ZZ, Yu H, Luo JY, et al., 2024. Automate strategy finding with LLM in quant investment. https://doi.org/10.48550/arXiv.2409.06289

Li YT, Yang X, Yang X, et al., 2025. R&D-Agent-Quant: a multi-agent framework for data-centric factors and model joint optimization. https://doi.org/10.48550/arXiv.2505.15155

Li ZW, Song R, Sun CH, et al., 2024. Can large language models mine interpretable financial factors more effectively? A neural-symbolic factor mining agent model. Findings of the Association for Computational Linguistics, p.3891-3902. https://doi.org/10.18653/v1/2024.findings-acl.233

Mehra S, Louka R, Zhang YX, 2022. ESGBERT: language model to help with classification tasks related to companies' environmental, social, and governance practices. https://doi.org/10.48550/arXiv.2203.16788

Mirjalili S, 2019. Genetic algorithm. In: Mirjalili S (Ed.), Evolutionary Algorithms and Neural Networks: Theory and Applications. Springer, Cham, p.43-55. https://doi.org/10.1007/978-3-319-93025-1_4

Nie YQ, Kong YX, Dong XW, et al., 2024. A survey of large language models for financial applications: progress, prospects and challenges. https://doi.org/10.48550/arXiv.2406.11903

OpenAI, 2023. GPT-4 technical report. https://doi.org/10.48550/arXiv.2303.08774

Papasotiriou K, Sood S, Reynolds S, et al., 2024. AI in investment analysis: LLMs for equity stock ratings. Proc 5th ACM Int Conf on AI in Finance, p.419-427. https://doi.org/10.1145/3677052.3698694

Real E, Liang C, So D, et al., 2020. AutoML-Zero: evolving machine learning algorithms from scratch. Proc 37th Int Conf on Machine Learning, p.8007-8019. https://doi.org/10.48550/arXiv.2003.03384

Shi H, Song WL, Zhang XT, et al., 2025. AlphaForge: a framework to mine and dynamically combine formulaic alpha factors. Proc 39th AAAI Conf on Artificial Intelligence, p.12524-12532. https://doi.org/10.1609/aaai.v39i12.33365

Shi Y, Duan YT, Li J, 2025. Navigating the alpha jungle: an LLM-Powered MCTS framework for formulaic factor mining. https://doi.org/10.48550/arXiv.2505.11122

Srivastava P, Malik M, Gupta V, et al., 2024. Evaluating LLMs' mathematical reasoning in financial document question answering. https://doi.org/10.48550/arXiv.2402.11194

Su HY, Wu K, Huang YH, et al., 2024. NumLLM: numeric-sensitive large language model for Chinese finance. https://doi.org/10.48550/arXiv.2405.00566

Tang ZY, Chen ZC, Yang JR, et al., 2025. AlphaAgent: LLM-driven alpha mining with regularized exploration to counteract alpha decay. https://doi.org/10.48550/arXiv.2502.16789

Wang SZ, Yuan H, Zhou L, et al., 2023. Alpha-GPT: human-AI interactive alpha mining for quantitative investment. https://doi.org/10.48550/arXiv.2308.00016

Wang SZ, Yuan H, Ni LM, et al., 2024. QuantAgent: seeking holy grail in trading by self-improving large language model. https://doi.org/10.48550/arXiv.2402.03755

Wang YN, Zhao JM, Lawryshyn Y, 2024. GPT-signal: generative AI for semi-automated feature engineering in the alpha research process. https://doi.org/10.48550/arXiv.2410.18448

Weng LL, 2023. LLM Powered Autonomous Agents. Lil'Log. https://lilianweng.github.io/posts/2023-06-23-agent [Accessed on June 21, 2025].

Wu SJ, Irsoy O, Lu S, et al., 2023. BloomberGPT: a large language model for finance. https://arxiv.org/abs/2303.17564

Xia L, Yang MM, Liu Q, 2024. Using pre-trained language model for accurate ESG prediction. Proc 8th Financial Technology and Natural Language and Proc 1st Agent AI for Scenario Planning, p.1-22. https://aclanthology.org/2024.finnlp-2.1

Yang X, Chen HT, Feng WJ, et al., 2024. Collaborative evolving strategy for automatic data-centric development. https://doi.org/10.48550/arXiv.2407.18690

Yu S, Xue HY, Ao X, et al., 2023. Generating synergistic formulaic alpha collections via reinforcement learning. Proc 29th ACM SIGKDD Conf on Knowledge Discovery and Data Mining, p.5476-5486. https://doi.org/10.1145/3580305.3599831

Yu YY, Yao ZY, Li HH, et al., 2024. FinCon: a synthesized LLM multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. Proc 38th Int Conf on Neural Information Processing Systems, Article 4354.

Yuan H, Wang SZ, Guo J, 2024. Alpha-GPT 2.0: human-in-the-loop AI for quantitative investment. https://doi.org/10.48550/arXiv.2402.09746

Zhang Q, Qin C, Zhang Y, et al., 2022. Transformer-based attention network for stock movement prediction. *Expert Syst Appl*, 202:117239. https://doi.org/10.1016/j.eswa.2022.117239

Zhang TP, Zhang ZYA, Fan ZY, et al., 2023. OpenFE: automated feature generation with expert-level performance. Proc 40th Int Conf on Machine Learning, p.41880-41901.