Vol. 16 (2026): Proceedings of the 2026 International Conference on AI and Technological Innovation Applications (AITIA 2026)
E-ISSN: 3105-0522 P-ISSN: 3105-0514
DOI: https://doi.org/10.70267/aitia.2026395403

From Traditional Methods to Large Language Models: A Comparative Study of Three Stock Return Prediction Models

Ruotong Chen*

University of South-Central Minzu, Wuhan, China

**Corresponding author: Ruotong Chen.*

Abstract

Stock return forecasting has shifted direction several times over the past thirty years, starting with linear factor models, moving through traditional machine learning, and lately turning to large language models that extract signals from text. Each approach has its own empirical literature, yet the three have rarely been compared within a single analytical frame. This paper takes on that comparison. Using the Fama-French model to represent the linear factor tradition, LASSO with random forests and neural nets for machine learning, and LLM-based text methods for the newest wave, we examine them across their theoretical foundations, the data they consume, how they model returns, and how they perform empirically. The findings are noteworthy. LLMs have, for the first time, brought unstructured text into the forecasting pipeline, and their predictive edge survives after standard factors are accounted for. But considerable unknowns remain. Long horizon robustness is unproven, hallucinations pose real risks, and cross-market portability is largely untested. The paper ends by sketching an integrated asset pricing framework, a conceptual starting point for braiding the three traditions together.

Keywords

stock return predictability, machine learning, large language models, asset pricing, text-driven forecasting

1. Introduction

Asset pricing is a central topic in financial economics. From the perspective of methodological evolution, stock return forecasting has undergone three significant phases: linear factor methods represented by the Fama-French three-factor [1] and five-factor models [2]; machine learning methods pioneered by Gu, Kelly, and Xiu [3]; and, in recent years, text analysis methods driven by large language models. Linear factor models have long dominated asset pricing research due to their high economic interpretability; however, their strict linear assumptions struggle to capture the widespread nonlinear relationships in the market (such as factor interaction effects, threshold effects, and time-varying risk premiums). Traditional machine learning methods have reduced prediction errors by more than 30% compared to linear models by automatically discovering complex nonlinear patterns [3], yet they face risks of overfitting [4] and “black-box” interpretability issues. The emergence of large language models (LLMs) marks a fundamental breakthrough: unlike the first two

categories of methods, which rely solely on structured numerical data, LLMs can extract semantic signals and sentiment trends from unstructured text, such as news articles, corporate announcements, and social media—and convert textual information into quantifiable predictive factors [5-6].

Although the three aforementioned methods have developed independently, few studies have systematically compared them within a unified framework. Of particular note is that the controversy surrounding the long-term robustness of LLMs has not yet been incorporated into a three-tiered comparative analytical framework [7]. Based on these findings, this paper proposes four core research questions: (1) What are the fundamental differences among the three paradigms in terms of core concepts, data sources, and interpretability? (2) Do large language models possess capabilities that the first two categories of methods cannot achieve? (3) What controversies and gaps still exist in current LLM asset pricing research? (4) How should the three paradigms be integrated in the future to achieve more efficient asset pricing? This paper adopts a representative case comparison method, selecting the most representative research findings from each approach to conduct a systematic comparison across four dimensions: theoretical foundations, data characteristics, modeling logic, and empirical performance.

2. Theoretical Foundations

2.1 The Efficient Market Hypothesis and Linear Factor Models

The Efficient Market Hypothesis serves as the logical starting point for understanding asset pricing theory [8]. Under the efficient market framework, stock returns are driven solely by systematic risk factors, and any forecasting strategy based on publicly available information cannot, in theory, generate risk-adjusted excess returns. The theoretical foundation of the linear factor model is rooted in arbitrage pricing theory [9], with its core assumption being that the expected return on an asset can be explained by a linear combination of common risk factors. The Capital Asset Pricing Model (CAPM), proposed by Sharpe and Lintner, treats market excess returns as the sole systematic factor and represents the simplest form of a linear model. Building on this, Fama and French introduced the size factor (SMB) and the high book-to-market ratio factor (HML) to construct a three-factor model [1]. The theoretical rationale is that small-cap companies face higher operational and liquidity risks, while companies with high book-to-market ratios (value stocks) reflect the risk of financial distress, both of which require corresponding premium compensation. Subsequently, Fama and French further introduced the profitability factor (RMW) and the capital maintenance factor (CMA), forming a five-factor model [2] in response to the empirical findings on profitability and investment effects by Titman et al. and Novy-Marx [10].

The core advantage of the linear factor model lies in its clear economic logic: each factor corresponds to a distinct source of risk, and the predictive results can be directly explained through factor loadings and risk premiums. However, its fundamental limitation lies in the linear assumption itself. When the actual data generation process involves nonlinear structures such as asymmetric effects and state dependence, linear approximation will inevitably lead to model specification bias. Factor construction methods deserve scrutiny for another reason. Grouping stocks by market cap and book-to-market ratio involves researcher discretion, and different grouping schemes can shift empirical conclusions in nontrivial ways [11]. McLean and Pontiff [12] showed that the predictive power of published anomaly factors drops by more than 20 percent on average in out-of-sample tests. The natural reading is that investors learn, and trading on a public factor gradually prices it away.

2.2 The Theoretical Rationale and Methodological Challenges of Machine Learning

Machine learning marked a shift for asset pricing. The field had long worked from theory to data, specifying functional forms that economic logic demanded. Machine learning flipped the direction, letting patterns emerge from the data itself.

The intellectual roots lie in Vapnik's statistical learning theory. Regularization disciplines model complexity when samples are large enough, which helps navigate the bias-variance tradeoff without tipping into overfit territory. In practice, tree-based methods like random forests and gradient boosting naturally capture interaction effects and piecewise structures that linear models miss. Neural networks take a different route, stacking layers of nonlinear transformation to approximate nearly any functional shape.

Gu, Kelly, and Xiu [3] put these ideas to a large-scale test. Using monthly returns on roughly 30,000 U.S. stocks from 1957 to 2016, they ran thirteen methods head-to-head. Two results stood out. First, nonlinear models beat linear benchmarks decisively, with shallow neural networks leading the pack. The out-of-sample R^2 climbed from 0.26 under a linear model to 0.40, a gain exceeding 50 percent. Second, price momentum was the single strongest predictor in every method, while fundamental variables of the sort central to Fama-French logic contributed surprisingly little. That gap is telling.

But predictive accuracy tells only part of the story. Kozak, Nagel, and Santosh [4] argued that the deeper problem for machine learning in asset pricing is not weak fit but missing economic discipline. Purely statistical models can latch onto in-sample patterns that carry no real economic rationale, and such patterns rarely hold up out of sample. They proposed injecting economic theory back in through Bayesian priors shaped by mean-variance efficient portfolios. Whether this, or something like it, can supply the missing discipline remains an open question.

2.3 The Technical Foundation and Financial Application Logic of Large Language Models

The emergence of large language models (LLMs) marks a qualitative shift in the information frontier of asset pricing. The first two families of methods both depend on structured numerical data. LLMs depart from this in a fundamental way. They can digest and interpret unstructured natural language, and that is where their real novelty lies.

On the technical side, modern LLMs rest on the Transformer architecture. The self-attention mechanism lets the model dynamically weigh how each word in a text relates to every other word [13], enabling it to track dependencies across long passages. The standard recipe runs unsupervised pre-training on vast corpora of general text first [14-15], then supervised fine-tuning on narrower tasks, financial sentiment analysis being an obvious example.

The economic logic traces back to an early finding by Tetlock. He showed that when pessimistic language piles up in the media, stock prices tend to come under pressure [16], mapping a rough pathway from text through sentiment to price. Loughran and McDonald [17] complicated this picture. Financial language, they noted, carries meanings that off-the-shelf sentiment tools routinely miss. A word that reads as neutral in everyday usage can signal something quite different in a 10-K filing. LLMs handle this mismatch more naturally because they work from context rather than keyword matching.

LLMs are not best understood as a replacement for older methods. They represent an expansion along the information frontier. Asset pricing used to run on quantified numbers. Now it can also draw on language, the medium through which market participants voice expectations, fears, and beliefs. These are forces that traditional models could never observe head-on, even though they shape prices every day.

3. Literature Review and Analysis

3.1 Linear Factor Models: Empirical Evidence and Structural Limitations

Empirical research on linear factor models has accumulated over more than three decades. Based on data from the U.S. market from 1963 to 1990, Fama and French demonstrated that a three-factor model can explain the vast majority of cross-sectional return variance [1], with explanatory power significantly superior to that of the CAPM. Cross-country validation studies indicate that the scale effect and value premium are prevalent across major global markets [18], though the magnitude of these effects and their statistical significance vary by region. Fama and French's five-factor model further enhances explanatory power; however, the authors themselves acknowledge that, following the introduction of the profitability and investment factors, the HML factor becomes redundant in many markets [2], suggesting that information overlap among factors may have been underestimated.

From a methodological perspective, the limitations of the linear factor model manifest on three levels. First, the arbitrary nature of variable selection. The composition of the model's factors is entirely defined in advance by the researcher, and any variables not theoretically presumed to be important will be systematically excluded. For example, irrational factors such as investor sentiment and media attention are difficult to incorporate into the linear risk premium framework. Second, the arbitrariness of factor construction. The grouping scheme for

factors (e.g., 5×5 or 2×3 groupings) has a substantial impact on the results [11], yet there is no theoretical consensus on the optimal grouping scheme. Third, the systemic risk of factor failure. McLean and Pontiff's empirical analysis shows that the average return of anomaly factors from 97 academic papers declined by an average of 58% after publication, with approximately 32 percentage points attributable to investor learning effects, with the decline concentrated after the factors were publicly disclosed [12], strongly suggesting that investor learning effects are systematically eroding the factors' pricing power. However, it should be noted that the high interpretability of linear factor models endows them with a unique “error-correction” capability: since the economic implications of each factor are clear, researchers can distinguish genuine risk factors from pseudo-factors derived from data mining through theoretical analysis and out-of-sample testing, and this advantage is a relative weakness of machine learning and LLM methods.

3.2 Traditional Machine Learning: Breakthroughs in Predictive Accuracy and the Cost of Interpretability

The work of Gu, Kelly, and Xiu [3] has been central to bringing machine learning into the asset pricing conversation. Two patterns in their results deserve particular attention. The first is an inverted U-shaped relationship between model complexity and predictive accuracy. Shallow neural networks with two or three hidden layers consistently beat deeper architectures, which suggests that the signal-to-noise ratio in financial data may simply be too low to sustain highly complex models. Beyond a certain point, adding more layers fits noise rather than genuine structure. The second pattern concerns what actually drives predictions. Price momentum turned out to be the strongest predictor across every method they tested, eclipsing the fundamental variables that anchor the Fama-French framework. This is telling. The pricing structure that machine learning uncovers looks systematically different from the one that decades of financial theory have built up from first principles, and that gap raises questions that have not yet been fully reckoned with.

Seen methodologically, machine learning faces three stubborn problems in asset pricing. One is overfitting and out-of-sample robustness. Kozak et al. [4] note that with limited samples, machine learning models tend to mistake noise for signal, and finance happens to be a setting where this risk is especially severe. The signal-to-noise ratio here is far worse than in image recognition or NLP. Another is the nagging lack of economic interpretability. The black box nature of these models means their predictions resist decomposition into economically meaningful pieces. An investor needs to know why, not just what, and machine learning seldom delivers an answer that can be acted on with confidence. The third problem concerns the information frontier itself. Machine learning can flexibly model complex relationships among numerical variables, but its inputs remain confined to structured data. A great deal of textual material, earnings call transcripts, the MD&A of financial reports and analyst revisions, the sentiment threaded through news coverage remains largely untapped by the first two families of methods.

3.3 Text-Based Prediction Methods Using Large Language Models: Cutting-Edge Advances Amid Controversy

The application of large language models (LLMs) in asset pricing represents a key leap in the information dimension from “numbers” to “text”, and recent research has been characterized by both rapid development and deep controversy.

Dong, Stratopoulos, and Wang [5] reviewed roughly 260 publications and identified several domains where ChatGPT and similar LLMs are gaining ground in finance. Automated text analysis, spanning sentiment classification and information extraction, is one. Investment decision support that includes factor generation and return forecasting is another, alongside audit compliance detection and academic research assistance. A clear pattern is that LLMs outperform traditional lexicon tools like the Loughran-McDonald dictionary and earlier methods like SVM on financial sentiment tasks, and by a meaningful margin. The edge comes from context. Revenue growth of 20 percent is good news. Revenue growth falling from 30 to 20 percent is not. Yet traditional lexicon methods tend to score both as positive. LLMs separate the two because they track how words relate to each other, not what they mean in isolation. Wang, Sun, and Wang [6] paired LLM sentiment extraction with a selective state-space model to push this line of work forward. Their empirical results show that LLM-based news sentiment factors can still generate significant excess returns (α) of approximately 4%–7% annually even after controlling for the Fama-French five-factor model and the momentum factor; however, the predictive power of this signal is highly concentrated in the short term (1 day to 1 week) and rapidly

declines in the medium to long term (1 month or longer), which suggests that the LLM signal may capture more of the market's short-term reaction bias to information rather than long-term risk premiums.

Cheng and Tang approached the problem from the perspective of factor generation, allowing GPT-4 to autonomously read academic literature, propose potential pricing factors, and conduct backtesting [19]. The results showed that the factor combinations generated by GPT-4 could explain a portion of the excess returns not captured by the five-factor model, and some factors had sound economic rationale (such as the supply chain concentration factor). However, critics point out that GPT-4's training data may already have included relevant information from the literature. If the model is merely "recalling" rather than "innovating", the originality of the factors it generates would be significantly diminished.

Darwish, Hassanien, and Eissa provide the most balanced critical review to date, systematically outlining five core controversies [7]: (1) Barriers to data accessibility: high-quality financial text data often requires expensive subscriptions, and LLM service providers do not disclose the sources of their training data, severely limiting the reproducibility of research; (2) Doubts about long-term robustness: the backtesting windows in existing studies span only 3–5 years, which is insufficient to evaluate model performance during extreme events such as financial crises or liquidity crunches; (3) Hallucination risk: LLMs may generate analyses that appear plausible on the surface but are substantively incorrect. In financial contexts, hallucinations manifest in three specific forms: (i) Temporal hallucination. Misattributing the timing of corporate events, which can invert the direction of event-study returns; (ii) Entity hallucination. Conflating information from companies with similar names, tickers, or industry classifications, a risk amplified by the prevalence of abbreviated company names in financial news; (iii) Numerical hallucination. Fabricating plausible-sounding financial figures (e.g., revenue growth rates) that do not exist in any source document, which is especially dangerous because generated numbers often fall within realistic ranges and may pass superficial plausibility checks; (4) Lack of evaluation standards: Differences across studies in model selection, prompt design, and benchmarking make cross-comparisons nearly impossible; (5) Insufficient real-time adaptability: LLM model update cycles are measured in months or years, making it difficult to respond promptly to sudden market events.

Beyond the five controversies enumerated above, a deeper methodological concern warrants attention: training data leakage. Unlike traditional models where training and evaluation datasets are clearly delineated by temporal splits, LLMs are pre-trained on web-scale corpora whose exact composition and temporal coverage are opaque. Critically, these training corpora may, through web crawls, include financial news articles, analyst reports, and even academic papers describing stock return anomalies that were published before the model's knowledge cutoff but after the prediction dates in backtesting studies. If an LLM's "prediction" of a factor's effectiveness is partially driven by having encountered the factor's published backtest results in its training data, the apparent predictive power conflates genuine financial reasoning with ex-post memorization. This concern is particularly acute for studies such as Cheng and Tang [19], where GPT-4 "autonomously" proposes pricing factors by reading academic literature, and if the model's training corpus included the very papers documenting those factors' historical performance, the distinction between innovation and recall becomes difficult to adjudicate. A rigorous mitigation strategy would require temporal holdout protocols: LLM-based factor generation or sentiment extraction must use only information that was publicly available before the prediction window, and the LLM's knowledge cutoff must be verified against the prediction dates. So far, no published LLM-in-finance study has implemented such a protocol.

4. Key Findings

4.1 Comprehensive Insights

After placing the three categories of methods within a unified comparative framework, the following core conclusions deserve special emphasis.

First, the three methods are not simply substitutes for one another but rather progress and evolve along two axes: "expansion of information dimensions" and "enhancement of modeling flexibility". Linear factor models maintain an absolute advantage in the "interpretability" dimension; traditional machine learning leads in "pattern discovery within structured data"; and LLMs have achieved a groundbreaking leap in the "ability to utilize unstructured information", the strategic significance of which lies in the fact that the information frontier

of asset pricing has been expanded from financial figures to linguistic text, and language is precisely the core medium through which market participants express expectations, convey sentiment, and form consensus.

Second, no single method can achieve optimal performance across all evaluation dimensions simultaneously. This implies that future research should focus not on "selecting the best model", but on "designing the optimal combination". Systematically integrating the economic interpretability of linear models, the nonlinear pattern recognition capabilities of machine learning, and the semantic understanding capabilities of LLMs is expected to generate synergies that surpass any single method.

Third, the most far-reaching potential impact of LLMs lies not in marginal improvements in predictive accuracy, but in the redefinition of asset pricing methodology itself. When models are capable of "reading" and "understanding" vast amounts of text, the concept of "information efficiency", the speed and adequacy with which prices reflect information—will take on entirely new connotations and dimensions of measurement. However, whether this potential can be translated into stable, reliable, and reproducible empirical results still depends on whether the academic community can properly resolve the five core controversies mentioned above. In a sense, LLMs are still in the "demonstration of capability" phase in the field of asset pricing and have not yet fully entered the "reliability verification" phase [7].

4.2 A Multidimensional Systematic Comparison of the Three Methodologies

Based on the aforementioned literature review, this section systematically compares the three forecasting paradigms across four core dimensions.

First, the data source dimension shows an evolutionary trend from "closed" to "open". Both linear factors and traditional machine learning rely on structured numerical data, and their common limitation lies in the fact that they "can only process quantified information". LLMs have broken through this boundary for the first time, incorporating unstructured text, such as news reports, financial statement narratives, analyst reports, and social media discussions, into the forecasting framework. This breakthrough enables a vast amount of qualitative information previously excluded from models, such as management tone, descriptions of the competitive landscape, and statements on regulatory policies, to be "translated" into quantifiable forecasting signals.

Second, the modeling logic has undergone a paradigm shift from "theory-driven" to "data-driven" and then to "language-driven". Linear factor models strictly adhere to the functional forms prescribed by economic theory; traditional machine learning relaxes the linear assumption but must search within a feature space predefined by researchers. LLMs, however, further automate the "definition of features" itself, for example, the model autonomously learns from text which semantic patterns are correlated with future returns.

Third, the dimension of interpretability is where the differences are most pronounced and the debates most intense. Linear factor models are inherently highly interpretable due to their clear economic logic; traditional machine learning sacrifices interpretability in exchange for predictive accuracy; LLMs occupy a unique middle ground, while the mechanisms of their billions of internal parameters far exceed the scope of direct human understanding, they can generate explanations in natural language (such as "The reason for the upward revision in the forecast is that quarterly revenue exceeded expectations"). However, the degree of consistency between these "narrative explanations" and their actual decision-making logic currently lacks rigorous methodological validation.

Finally, the dimension of limitations reveals a structural difference. The limitations of linear factor models (linear assumptions, factor failure) essentially stem from a lack of methodological "rigidity", the models are too simple to adapt to complex real-world situations. The limitations of traditional machine learning (overfitting and lack of interpretability) essentially stem from excessive "flexibility", they may capture noise rather than signals; whereas the limitations of LLMs (hallucinations, high cost, and lack of stability) reflect a more fundamental uncertainty. Although the models are powerful, their behavioral boundaries and failure modes have not yet been fully understood.

Table 1: Four-Dimension Systematic Comparison of the Three Forecasting Paradigms.

	Linear Factor Models	Traditional Machine Learning	LLM-Based Text Pre-diction
Data Sources	Structured numerical data (firm characteristics, macro variables); closed set of pre-specified factors	Structured numerical data; flexible but limited to quantified inputs	Structured numerical data + unstructured text (news, filings, social media); open-ended information frontier
Modeling Logic	Theory-driven; functional form prescribed by economic theory (arbitrage pricing, linear risk premia)	Data-driven; functional form learned from data (trees, NNs); regularization controls complexity	Language-driven; semantic patterns extracted from text; LLMs partially automate feature definition itself
Interpretability	High: each factor maps to an economic risk source; factor loadings directly interpretable	Low: black-box predictions; SHAP/LIME provide post-hoc approximations without structural economic decomposition	Mixed: internal mechanisms opaque (billions of params), but can generate natural-language explanations whose fidelity is unverified
Key Limitations	Linearity assumption; factor construction arbitrariness; post-publication decay (58% average return decline)	Overfitting risk; noise-signal confusion in low-SNR financial data; lacks economic constraint mechanisms	Hallucination (temporal, entity, numerical); data leakage from training corpus; prompt sensitivity harms replicability

5. Comprehensive Discussion

5.1 What the Findings Mean for Practice

For quantitative investors, the practical takeaway is that LLMs work best as complements to existing models, not as replacements. A three tiers information architecture is one plausible design. Linear factors anchor the baseline risk model, machine learning captures high dimensional nonlinear pricing patterns, and LLM text signals provide a stream of incremental alpha. Since LLM signals tend to be short lived, they are likely better suited to high frequency or intraday strategies than to long horizon asset allocation [6].

Regulators face a somewhat different set of questions. Maliciously exploited hallucinations could amount to a new strain of market manipulation. The concentration of LLM services among a few commercial providers also risks widening the information gap between institutional and retail investors. Two responses seem worth weighing. Requiring disclosure when investment analysis draws on LLM generated output is one. Supporting the development of open source financial LLMs to lower entry barriers is another.

5.2 Avenues for Future Work

This paper also offers insights for future research. It proposes a conceptual framework for "Integrated Asset Pricing", which comprises three core principles: (1) Information Integration: the system fuses structured numerical data with unstructured text data; (2) Methodological Integration: organically combining the economic constraints of linear models, the pattern-discovery capabilities of machine learning, and the semantic understanding capabilities of LLMs; for example, constructing a hybrid workflow such as "LLM text feature extraction → LASSO dimensionality reduction and regularization → LSTM time-series modeling"; (3) Evaluation Integration: establishing a multi-objective evaluation system covering predictive accuracy, economic value, interpretability, and robustness. In terms of specific research directions, cross-market and cross-language validation, hybrid model architecture design, and methodological breakthroughs in online learning adaptability and interpretability are the four priority topics.

A worked example would present as follow. A concrete hybrid workflow following this architecture is: Step 1—A fine-tuned BERT model classifies the sentiment of management discussion and analysis (MD&A) sections from 10-K filings, producing a monthly sentiment score per firm. Step 2—LASSO regression selects, from a pool of 94 firm characteristics plus the BERT sentiment score, the subset of predictors with non-zero coefficients, imposing sparsity to mitigate overfitting. Step 3—An LSTM network models the temporal dynamics of the selected predictors to produce monthly return forecasts. Step 4—SHAP values decompose each forecast into the marginal contribution of each input feature, mapping predictions back to their economic drivers (e.g., "30% of the positive forecast is attributable to improving MD&A sentiment, 25% to momentum,

and 20% to value characteristics"). This modular pipeline illustrates how LLM signals complement rather than replace traditional factors, with the interpretability layer providing an audit trail for investment decisions.

6. Conclusion

By systematically reviewing the theoretical foundations and empirical literature on linear factor models, traditional machine learning, and text prediction methods based on large language models in the field of asset pricing, this paper establishes a three-tiered comparative analysis framework and draws the following main conclusions.

First, these three approaches constitute a methodological evolution ranging from "theory-driven linear simplicity" to "data-driven nonlinear discovery" and finally to "language-driven semantic understanding". Linear factor models have long dominated the field due to their high economic interpretability and concise computational logic; however, strict linear assumptions and reliance on structured data have limited further improvements in predictive capability. Traditional machine learning achieved substantial breakthroughs in predictive accuracy by introducing nonlinear modeling, but the costs, including the risk of overfitting and "black-box" lack of interpretability, are equally significant; large language models (LLMs) have, for the first time, incorporated unstructured textual information into the predictive framework, demonstrating significant incremental predictive power even after controlling for traditional factors, yet key controversies, such as insufficient long-term robustness, the risk of hallucinations, and high costs, still remain unresolved.

Second, the core contributions of this paper are as follows: (1) Within the scope of our literature search, this review is among the earliest to incorporate three categories of asset pricing methods into a systematic three-tier comparative framework, conducting a structured analysis across four dimensions: data sources, modeling logic, interpretability, and limitations; (2) Through "comprehensive insights", it reveals that the three methods are not mutually exclusive but each possess comparative advantages across different evaluation dimensions, this finding lays the logical foundation for the "integrated asset pricing" framework; (3) Building on Darwish et al.'s [7] recent framework, clearly identifying the five key conditions required for LLMs to transition from the "demonstration of capability" stage to the "verification of reliability" stage.

References

[1] Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. Journal of Financial Economics, 33(1), 3–56. https://doi.org/10.1016/0304-405X(93)90023-5

[2] Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. Journal of Financial Economics, 116(1), 1–22. https://doi.org/10.1016/j.jfineco.2014.10.010

[3] Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. The Review of Financial Studies, 33(5), 2223–2273. https://doi.org/10.1093/rfs/hhaa009

[4] Kozak, S., Nagel, S., & Santosh, S. (2020). Shrinking the cross-section. Journal of Financial Economics, 135(2), 271–292. https://doi.org/10.1016/j.jfineco.2019.06.008

[5] Dong, M. M., Stratopoulos, T. C., & Wang, V. X. (2024). A scoping review of ChatGPT research in accounting and finance. International Journal of Accounting Information Systems, 55, 100715. https://doi.org/10.1016/j.accinf.2024.100715

[6] Wang, R., Sun, M., & Wang, L. (2025). From news to trends: A financial time series forecasting framework with LLM-driven news sentiment analysis and selective state spaces. Journal of Intelligent Information Systems, 63(6), 1955–1980. https://doi.org/10.1007/s10844-025-00971-3

[7] Darwish, M., Hassanien, E. E., & Eissa, A. H. B. (2026). Stock Market Forecasting: From Traditional Predictive Models to Large Language Models. Computational Economics, 67(6), 4553–4597. https://doi.org/10.1007/s10614-025-11024-w

[8] Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. Journal of Finance, 25(2), 383–417. https://doi.org/10.2307/2325486

[9] Ross, S. A. (1976). The arbitrage theory of capital asset pricing. Journal of Economic Theory, 13(3), 341–360. https://doi.org/10.1016/0022-0531(76)90046-6

[10] Novy-Marx, R. (2013). The other side of value: The gross profitability premium. Journal of Financial Economics, 108(1), 1–28. https://doi.org/10.1016/j.jfineco.2013.01.003

[11] Lo, A. W., & MacKinlay, A. C. (1990). Data-snooping biases in tests of financial asset pricing models. Review of Financial Studies, 3(3), 431–467. https://doi.org/10.1093/rfs/3.3.431

[12] McLean, R. D., & Pontiff, J. (2016). Does academic research destroy stock return predictability? Journal of Finance, 71(1), 5–32. https://doi.org/10.1111/jofi.12365

[13] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (Vol. 30, pp. 5998–6008). Curran Associates.

[14] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423

[15] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., … & Amodei, D. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems (Vol. 33, pp. 1877–1901). Curran Associates.

[16] Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. Journal of Finance, 62(3), 1139–1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x

[17] Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. Journal of Finance, 66(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

[18] Fama, E. F., & French, K. R. (2012). Size, value, and momentum in international stock returns. Journal of Financial Economics, 105(3), 457–472. https://doi.org/10.1016/j.jfineco.2012.05.011

[19] Cheng, Y., & Tang, K. (2024). GPT's idea of stock factors. Quantitative Finance, 24(9), 1301–1326. https://doi.org/10.1080/14697688.2024.2318220

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Open Access