Levesque Marie
September 12, 2025
MSc Data Science, University of Essex Online

The Role of Large Language Models in Financial Market Sentiment and Investment Research

1. Introduction

The role of sentiment in shaping financial markets has been widely recognised for decades, with investor emotions influencing asset prices, risk perceptions, and trading behaviour. Traditional quantitative models have long attempted to capture sentiment signals from structured and unstructured data, particularly through text-based sources such as news articles, analyst reports, and social media. However, early sentiment approaches, such as lexicon-based dictionaries and basic machine learning models, often struggled to capture the context-dependent and domain-specific nuances of financial language. This limitation has spurred growing interest in advanced natural language processing (NLP) methods and, more recently, large language models (LLMs).

LLMs, including general-purpose models such as GPT-3 and GPT-4, and domain-specific adaptations like BloombergGPT and FinBERT, have transformed the landscape of financial text analysis. These models demonstrate a remarkable ability to extract, interpret, and contextualise sentiment from diverse sources of financial information at scale (Bloomberg, 2023). Empirical evidence suggests that their integration into investment research pipelines has the potential to enhance predictive accuracy and improve portfolio decision-making (Frontiers in Artificial Intelligence, 2025).

Nevertheless, the rapid adoption of LLMs within financial services also raises critical concerns. Regulators and policymakers have highlighted risks relating to transparency, fairness, bias, and the environmental costs of training large-scale models (European Securities and Markets Authority and The Alan Turing Institute, 2025). These considerations underline the importance of critically evaluating the opportunities and limitations of LLMs for sentiment-driven investment research.

This literature review therefore assesses current knowledge in this area, synthesising evidence from academic and industry sources to highlight both the promise and the challenges of applying LLMs to financial sentiment analysis and investment research.

2. Foundations of Sentiment Analysis in Finance

The use of sentiment analysis in finance predates the rise of large language models and is rooted in efforts to quantify qualitative information for trading and investment decisions. Early methods were predominantly lexicon-based, relying on predefined dictionaries of positive and negative words to classify the tone of financial texts. One of the most widely adopted examples is the Loughran–McDonald financial sentiment dictionary, which adapted general-purpose lexicons for financial contexts. Although these approaches

provided a systematic method to gauge investor sentiment, they often struggled with contextual ambiguity, domain-specific terminology, and sarcasm or negation in text. For instance, terms such as “liability” or “risk” may carry neutral or technical meanings in financial disclosures but would be incorrectly classified as negative in general lexicons. The limitations of lexicon-based approaches led to the adoption of machine learning techniques that could learn sentiment classifications from labelled data. Traditional supervised models such as support vector machines, logistic regression, and random forests improved upon dictionary methods by incorporating statistical patterns. With the rise of deep learning, recurrent neural networks (RNNs) and long short-term memory networks (LSTMs) became popular for their ability to capture sequential dependencies in text, while convolutional neural networks (CNNs) demonstrated effectiveness in identifying local sentiment features. Applications extended across multiple financial domains, including analysing earnings call transcripts, classifying analyst reports, and extracting market sentiment from news headlines and social media streams.

Despite these advances, pre-LLM machine learning models were not without significant drawbacks. They typically required substantial amounts of annotated training data, which is costly and time-consuming to produce in the financial domain. Furthermore, these models often failed to generalise across contexts, with performance dropping when applied to new asset classes, sectors, or time periods. Domain-specific adaptations, such as FinBERT, marked an important step forward by fine-tuning transformer-based architectures on financial text corpora (Araci, 2019). While FinBERT improved context sensitivity and interpretability compared with earlier approaches, it remained limited by corpus size and lacked the broader reasoning capacity of more recent LLMs.

Overall, the evolution from lexicon-based to deep learning methods highlights the progressive refinement of sentiment analysis in finance. However, the persistent challenges of domain specificity, limited generalisation, and costly annotation provided the motivation for exploring LLMs, which promise greater contextual understanding and transferability across diverse financial tasks.

3. LLMs in Financial Market Sentiment & Investment

The emergence of large language models has marked a paradigm shift in the application of natural language processing within financial markets. Unlike traditional sentiment analysis approaches, which often struggled with contextual nuances and domain-specific vocabulary, LLMs offer advanced contextual understanding, transfer learning, and scalability. This section critically examines their role in financial sentiment and investment research, considering both general-purpose and domain-specific models, as well as their demonstrated applications.

3.1. General-Purpose LLMs

Mainstream models such as GPT-3, GPT-4, Claude, and LLaMA have been widely explored in financial sentiment tasks, despite not being trained specifically on finance. Their strength lies in their ability to generalise across diverse contexts, making them useful for extracting sentiment from heterogeneous sources such as social media, news feeds, and policy announcements. Studies show that GPT-based models can outperform traditional lexicon-based and machine learning methods in zero-shot or few-shot settings (Zhang, Liu and Xu, 2025). For example, GPT-4 has demonstrated higher accuracy in sentiment classification of analyst reports without the need for additional fine-tuning (Wang, Li and Zhang, 2025).

However, reliance on general-purpose LLMs is not without challenges. Hallucinations, inconsistent outputs, and lack of domain precision remain concerns, particularly when such models encounter technical financial jargon or ambiguous phrasing. Moreover, as ESMA and the Alan Turing Institute (2025) caution, their opacity and difficulty of auditing pose significant risks in regulated contexts such as investment decision-making.

3.2. Domain-Specific LLMs

Recognising the limitations of general-purpose models, several domain-specific financial LLMs have emerged. FinBERT (Araci, 2019) was one of the earliest attempts, fine-tuning BERT on financial corpora to improve sentiment analysis. More recently, BloombergGPT (2023) was introduced as a 50-billion parameter model trained on 363 billion tokens of financial data, combining proprietary Bloomberg datasets with public sources. Benchmark results demonstrated its superior performance on finance-specific NLP tasks compared with both general LLMs and earlier domain models.

FinGPT represents another significant development, providing an open-source, cost-efficient alternative to BloombergGPT. With fine-tuning costs under $300, it aims to democratise access to financial LLMs while incorporating reinforcement learning from human feedback (RLHF) and retrieval-augmented generation (RAG) techniques. Similarly, EnhancedFinSentiBERT (ScienceDirect, 2025) illustrates the ongoing refinement of domain-specific architectures, reporting improved accuracy in sentiment detection for financial disclosures and news reports.

While these models demonstrate notable performance improvements, they also raise questions regarding accessibility and fairness. Proprietary models such as BloombergGPT remain closed and resource-intensive, creating barriers for smaller firms and academic researchers. By contrast, open-source initiatives such as FinGPT attempt to mitigate these inequalities, though their reliance on limited datasets may restrict robustness.

3.3. Applications in Investment Research

The primary application of LLMs in finance lies in their ability to analyse unstructured text data for investment insights. Three key areas have emerged in the literature:

News and Social Media Sentiment
The influence of news headlines and social media platforms such as Twitter and Reddit on market sentiment is well established. LLMs have demonstrated an ability to capture subtle shifts in tone more effectively than dictionary-based methods. Qin, Chen and Hu (2024) show that OPT, a GPT-3 derivative, achieved 74.4% accuracy in classifying financial news sentiment compared with ~50% for Loughran–McDonald dictionaries. When applied to trading strategies, the LLM-based models generated significantly higher Sharpe ratios, indicating economic relevance beyond mere classification accuracy.

Analyst Reports and Earnings Calls
Analyst notes and earnings call transcripts are critical sources of qualitative information for institutional investors. Kirtac and Germano (2025) demonstrate that integrating LLM-derived sentiment into trading strategies can substantially improve returns, with their self-financing portfolio achieving a Sharpe ratio of 3.05. MarketWatch (2025) similarly reports on a Yale study applying AI to 1.1 million analyst notes, which found that the textual content of these reports was more predictive of long-term performance than analysts' numerical price targets. These findings suggest that LLMs can distil hidden insights from narrative disclosures often overlooked by conventional analysis.

Macroeconomic and Policy Documents
Another emerging area is the use of LLMs to interpret central bank communications, regulatory announcements, and policy statements. Tang, Yang and Wang (2024) illustrate that instruction-tuned and RAG-enhanced LLMs achieved accuracy improvements of 1%–6% in classifying macroeconomic texts, which translated into higher portfolio Sharpe ratios. These results highlight the potential of LLMs to serve as tools for real-time macro sentiment monitoring, supporting risk management and asset allocation decisions.

3.4. Benchmarks and Comparative Evidence

The growing adoption of LLMs in finance has spurred the development of benchmarks tailored to the domain. FiQA, BloombergGPT benchmarks, and FinBench provide comparative evaluations of models on tasks such as question answering, summarisation, and sentiment classification. Results consistently show domain-specific LLMs outperforming general-purpose counterparts in specialised tasks (Bloomberg, 2023; ScienceDirect, 2025).

Nevertheless, findings remain mixed across studies. While Kirtac and Germano (2025) and Qin et al. (2024) report substantial economic gains from LLM-driven strategies, Wang, Li and Zhang (2025) argue that simpler zero-shot applications of GPT-4 often perform as well as or better than complex chain-of-thought prompting. This divergence suggests that

while some studies demonstrate strong performance, others highlight risks of overfitting and inconsistent results across datasets.

4. Critical Perspectives

While large language models offer notable advances in financial sentiment analysis, the literature also highlights several critical challenges that temper their practical value. These limitations relate to issues of accuracy, interpretability, bias, regulation, and broader systemic risks.

4.1. Accuracy and Reliability

Although empirical studies suggest that LLMs can outperform traditional models in many sentiment classification tasks, findings remain inconsistent. For instance, Kirtac and Germano (2025) report significant performance improvements, whereas Wang, Li and Zhang (2025) caution that reasoning prompts do not necessarily enhance accuracy, and in some cases zero-shot approaches perform equally well. These discrepancies raise concerns about reproducibility and overfitting, especially given the relatively short time horizons of many backtesting studies. The risk of hallucinations—outputs that are plausible but factually incorrect—further undermines the trustworthiness of LLM-based systems in high-stakes financial environments.

4.2. Bias and Data Limitations

Another pressing concern relates to bias and domain adaptation. LLMs trained on broad internet corpora may inadvertently inherit biases that distort financial sentiment analysis. Politically skewed or culturally specific language patterns could influence model outputs, potentially leading to systemic distortions in market signals. Domain-specific models such as BloombergGPT (2023) mitigate some of these risks by relying on financial data, but their proprietary and closed nature limits external validation. Open-source models like FinGPT attempt to democratise access, but their smaller training datasets may compromise robustness across asset classes and geographies.

4.3. Interpretability and Regulatory Challenges

Financial regulators emphasise that AI tools deployed in investment contexts must meet high standards of transparency and accountability. The ESMA and Alan Turing Institute (2025) highlight that the opacity of LLMs complicates auditing, making it difficult to explain or justify decisions derived from model outputs. This is particularly problematic given the requirements of the EU AI Act, which categorises financial decision-making tools as high-risk applications. Lack of interpretability not only raises compliance issues but also

undermines user confidence, especially in sectors such as asset management and insurance, where fiduciary responsibility is paramount.

4.4. Ethical and Systemic Risks

Finally, scholars and practitioners warn of broader ethical and systemic risks. Over-reliance on LLM-driven sentiment could amplify market herding behaviour, as many actors respond to similar AI-generated signals. Furthermore, the carbon footprint associated with training and deploying large models is increasingly scrutinised as a sustainability concern (ESMA and Alan Turing Institute, 2025). These risks suggest that LLMs should be viewed not as replacements for human judgment, but as tools that require careful oversight and integration into existing decision-making frameworks.

4.5. Summary of Section

Critical perspectives underscore that while LLMs may add value in investment research, their deployment must be accompanied by caution. Issues of bias, interpretability, regulation, and systemic risk highlight the necessity of combining LLM-driven insights with human expertise and rigorous governance mechanisms.

5. Conclusion

This review has examined the role of large language models in financial sentiment analysis and investment research, tracing developments from lexicon-based approaches through machine learning to state-of-the-art LLMs. Evidence indicates that LLMs significantly enhance the extraction and interpretation of sentiment from diverse sources, including financial news, analyst reports, and macroeconomic communications. Empirical studies highlight improvements in classification accuracy and economic outcomes, such as higher Sharpe ratios, underscoring their potential to support portfolio decision-making (Kirtac and Germano, 2025; Qin, Chen and Hu, 2024).

At the same time, critical perspectives emphasise important limitations. Concerns about bias, hallucinations, interpretability, and compliance illustrate that LLM adoption in finance cannot be uncritical (ESMA and Alan Turing Institute, 2025). The mixed results across empirical studies further suggest that expectations should be tempered.

Looking forward, the literature points to the need for hybrid approaches that combine LLMs with traditional quantitative models, as well as the development of standardised benchmarks for evaluation. Greater attention to governance, transparency, and sustainability will be necessary to ensure responsible adoption. In conclusion, LLMs represent a promising complement to traditional sentiment analysis techniques, but their future role will depend on careful integration with human expertise and robust oversight.

References

Araci, D. (2019) ‘FinBERT: Financial sentiment analysis with pre-trained language models’, *arXiv preprint*. Available at: https://arxiv.org/abs/1908.10063 (Accessed: 8 September 2025).

Bloomberg (2023) *BloombergGPT: A large language model for finance*. Available at: https://arxiv.org/abs/2303.17564 (Accessed: 12 September 2025).

European Securities and Markets Authority (ESMA) and The Alan Turing Institute (2025) *Large language models in finance*. Paris: ESMA. Available at: https://www.esma.europa.eu/sites/default/files/2025-06/LLMs_in_finance_-_ILB_ESMA_Turing_Report.pdf (Accessed: 12 September 2025).

Frontiers in Artificial Intelligence (2025) ‘Large language models in equity markets: applications, techniques, and insights’, *Frontiers in Artificial Intelligence*. Available at: https://www.frontiersin.org/articles/10.3389/frai.2025.1608365/full (Accessed: 12 September 2025).

Kirtac, K. and Germano, G. (2025) ‘Enhanced financial sentiment analysis and trading strategy development with large language models’, *SSRN Electronic Journal*. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5181105 (Accessed: 12 September 2025).

MarketWatch (2025) ‘This researcher put 1.1 million Wall Street analyst notes through AI. Here’s what it found’, *MarketWatch*. Available at: https://www.marketwatch.com/story/this-researcher-put-1-1-million-wall-street-analyst-notes-through-ai-heres-what-it-found-1b3e52d2 (Accessed: 12 September 2025).

Qin, Y., Chen, X. and Hu, Y. (2024) ‘Sentiment trading with large language models’, *arXiv preprint*. Available at: https://arxiv.org/abs/2412.19245 (Accessed: 12 September 2025).

ScienceDirect (2025) ‘EnhancedFinSentiBERT: A domain-adapted transformer for financial sentiment classification’, *Expert Systems with Applications*, 260, p. 125221. Available at: https://doi.org/10.1016/j.eswa.2025.125221 (Accessed: 12 September 2025).

Tang, H., Yang, Z. and Wang, J. (2024) ‘Aligning large language models with human instructions and stock market feedback for financial sentiment analysis’, *arXiv preprint*. Available at: https://arxiv.org/abs/2410.14926 (Accessed: 12 September 2025).

Taylor & Francis Online (2025) ‘Intraday stock prediction using sentiment analysis’, *Journal of Asset Management*. Available at: https://www.tandfonline.com/doi/full/10.1080/15427560.2025.2538879 (Accessed: 12 September 2025).

Wang, X., Li, Y. and Zhang, T. (2025) ‘Reasoning or overthinking: Evaluating LLMs on financial sentiment analysis’, *arXiv preprint*. Available at: https://arxiv.org/abs/2506.04574 (Accessed: 12 September 2025).

Zhang, Y., Liu, J. and Xu, K. (2025) ‘On assessing the performance of LLMs for target-level sentiment analysis in finance’, *Algorithms*, 18(1), 46. Available at: https://www.mdpi.com/1999-4893/18/1/46 (Accessed: 12 September 2025).