

A Multimodal Vision-Language Framework for Financial Anomaly Detection

Siang-Li JHENG*

*Marie Skłodowska-Curie Actions (MSCA) Digital Finance
Doctoral School of Cybernetics and Economic Statistics, Bucharest University of Economic
Studies, Romania*

IDA Institute Digital Assets

**Corresponding author, jhengsiang24@stud.ase.ro*

Daniel Traian PELE

*Bucharest University of Economic Studies, Bucharest, Romania
Institute for Economic Forecasting, Romanian Academy, Bucharest, Romania
IDA Institute Digital Assets
danpele@ase.ro*

Rahul TAK

*Marie Skłodowska-Curie Actions (MSCA) Digital Finance
Doctoral School of Cybernetics and Economic Statistics, Bucharest University of Economic
Studies, Romania;
IDA Institute Digital Assets
takrahul24@stud.ase.ro*

Stefan GAMAN

*Marie Skłodowska-Curie Actions (MSCA) Digital Finance
Doctoral School of Cybernetics and Economic Statistics, Bucharest University of Economic
Studies, Romania
Department of Econometrics, Poznań University of Economics and Business
IDA Institute Digital Assets
gamanstefan18@stud.ase.ro*

Abstract. Financial markets are complex systems where state estimation typically relies on either quantitative time-series models or subjective technical analysis. A gap remains in bridging scalable quantitative methods with the pattern-recognition intuition of human experts. In this work, we introduce a multimodal vision-language framework for financial anomaly detection that explores reframing market regime classification as a visual recognition task using advanced multimodal AI. We propose a methodology to project high-dimensional market indicators onto a standardized, multi-panel image format. We leverage the zero-shot capability of state-of-the-art Vision Language Models (VLMs), specifically GPT-4.1. We employ a rolling quantile-based approach for programmatic data labelling to reduce subjectivity. Furthermore, to address the need for transparency, our framework extracts structured natural language rationales directly from the model, offering interpretable insights alongside classifications. Our results, validated through out-of-sample testing on the S&P500, suggest that this multimodal prompting approach achieves promising fidelity in detecting systemic anomalies, indicating its potential as a reliable and interpretable framework for financial AI.

Keywords: Anomaly Detection, Large Language Models, Multimodal AI, Market Regime Classification, Time-Series Analysis.

DOI: 10.2478/picbe-2026-0020

© 2026 S.-L. Jheng; D. T. Pele; R. Tak, published by De Gruyter.

This work is licensed under the Creative Commons Attribution 4.0 License.

1 Introduction

An anomaly is fundamentally defined as a deviation from what is standard, normal, or expected. (Chandola et al., 2009) In the context of financial markets, these irregularities, ranging from sudden crashes to subtle regime shifts, represent critical turning points that can either wipe out years of gains or create significant opportunities for excess returns. Traditional risk management has long relied on statistical models such as GARCH (Bauwens et al., 2006) or Value-at-Risk (VaR) (Wipplinger et al., 2007) to detect these events. However, these classical methods typically operate under strict assumptions of stationarity and normality. As a result, they often fail to capture complex, non-linear black swan events, such as the 2008 financial crisis or the 2020 pandemic market crash, which defy standard statistical expectations.

While human analysts possess the intuition to spot these warning signs by synthesizing diverse information, reading news reports, interpreting technical chart patterns, and understanding market sentiment, this manual process is impossible to scale across thousands of assets on a daily basis. This limitation highlights a critical gap in current financial surveillance: the need for a system that combines the scalability of algorithms with the contextual understanding of a human expert. The emergence of Multimodal Large Language Models (MLLMs) offers a potential solution. Unlike previous generation models that processed only numerical data, these advanced systems can see charts and read context, effectively acting as a scalable team of junior analysts capable of identifying patterns that purely numerical models miss.

To bridge this gap, this paper introduces FINDER-LLM (FINancial Anomaly DETection and Recognition using Large Language Models), a novel and fully multimodal pipeline designed for financial anomaly detection and explanation. Moving beyond simple numerical forecasting, our approach integrates daily price data, technical indicators, visual chart images, and textual news headlines into a unified framework. Our work makes three primary contributions to the literature. First, we establish a comprehensive detection pipeline that utilizes Vision-Language Models to process heterogeneous market data simultaneously. Second, distinct from "black box" algorithms, our framework provides structured, analyst-oriented reasoning. It outputs not just an anomaly probability, but also a natural language rationale and key signals, explaining why a specific market movement is flagged as dangerous.

Third, we validate this methodology through a rigorous, reproducible benchmark. We compare the performance of Zero-shot prompting strategy against classical technical analysis baselines using strictly out-of-sample data from the S&P 500. By projecting high-dimensional market data into a format that LLMs can interpret visually and textually, FINDER-LLM demonstrates how AI can move from passive data processing to active, explainable risk recognition.

Overall, this study presents a comprehensive multimodal framework that connects quantitative market data with visual and textual analysis. The remainder of the paper is organized as follows: Section 2 reviews the literature on financial anomaly detection and the application of Vision-Language Models. Section 3 details the methodology of the FINDER-LLM framework, including data processing and prompting strategies. Section 4 presents the empirical results based on the S&P 500 index. Section 5 discusses the limitations of the current approach, and Section 6 concludes the study. Data, code, and regular updates for this paper are available at Quantlet ; the Quantinar  platform offers further media content for this research.

2 Literature review

The methodologies for identifying market irregularities have undergone a fundamental transformation, shifting from strict statistical frameworks to adaptive, data-driven intelligence. Historically, financial monitoring relied heavily on econometric models. Approaches such as GARCH and VaR became the industry standard for quantifying volatility and risk. While these models offer mathematical discipline, they are inherently limited by assumptions of stationarity and normality. Consequently, they frequently struggle to anticipate black swan events or complex regime shifts where market dynamics deviate sharply from historical distributions. The subsequent adoption of classical machine learning, including Support Vector Machines (SVM) (Cortes and Vapnik, 1995) and Random Forests (RF) (Breiman, 2001), addressed some of these linearity constraints but often failed to capture the semantic context behind market movements.

The advent of Large Language Models (LLMs) initially prompted researchers to treat financial time-series analysis as a natural language task. However, Zhou et al. (2024) demonstrated that LLMs exhibit a superior understanding of time-series data when it is presented as visual imagery rather than text. Their research found no evidence that explicit reasoning prompts via Chain-of-Thought (CoT) improve LLMs' performance in time series analysis. Furthermore, they noted that an LLM's understanding of anomalies is not consistent with human perception, and models tend to perform worse when the input time series has more tokens. This finding underscores a critical pivot in the field: the transition from purely textual or numerical processing to visual understanding.

This realization has led to a wave of research into Multimodal LLMs (MLLMs). Xu et al. (2025) introduced the VisualTimeAnomaly benchmark, establishing that MLLMs (e.g., GPT-4o, Gemini) can robustly detect time series anomalies by analyzing both chart images and contextual text. They found that few-shot prompts and hybrid inputs significantly boost detection accuracy, offering flexibility to new anomaly types, though they still lag behind classic machine learning in some specialized tasks. Further expanding on specific modalities, Park (2024) developed a multi-agent LLM system combining time-series and web text to validate anomalies in the S&P 500, while Yang et al. (2025) utilized LLMs to filter false positives from time series anomaly detectors using both charts and text. In other data formats, Tsai et al. (2025) introduced AnoLLM for LLM-based unsupervised anomaly detection on tabular data, and Liu et al. (2024) developed Time-MMD, a multiple-domain dataset showing multimodal gains in mean squared error (MSE). Similarly, Wang et al. (2023) proposed FinVis-GPT, a vision-language model fine-tuned on stock chart Q&A and trading signals using chart images and text. Finally, Zhang et al. (2025) expanded this domain by proposing CAMEF, which integrates text and time series to achieve causal forecasting from macro events.

Despite these rapid advancements, several distinct research gaps remain. First, there are no real-world multimodal benchmarks, as existing datasets lack aligned price, chart, and news data. Second, there is a single-modality focus in the field, most models use only price or text in isolation, thereby missing the critical cross-modal signals. Third, there is a severe lack of explainability, with very few systems offering analyst-friendly, structured outputs. Additionally, the literature exhibits poor robustness, as the handling of missing or noisy data is seldom studied. Finally, there are limited open solutions available, meaning that current pipelines are rarely fully open, reproducible, or modular.

3 Methodology

We propose FINDER-LLM, a framework designed to bridge the gap between quantitative financial econometrics and the adaptive reasoning of MLLMs. Unlike traditional supervised learning

approaches that require training specific architectures from scratch, our methodology leverages the pre-trained reasoning capabilities of state-of-the-art vision-language models, specifically GPT-4.1 Vision. Figure 1 shows the pipeline reframes anomaly detection as a multimodal inference task. As illustrated in our workflow, the FINDER-LLM framework employs a comprehensive pipeline. The system processes a sequence of rolling windows, integrating market data, textual data, external signals, and historical data. After data preprocessing, these inputs are converted into four aligned branches: technical analysis, NLP processing, visual encoding, and risk metrics. These modalities are synthesized by the Vision-Language Model to generate anomaly classifications, interpretable explanations, and confidence scores for downstream tasks such as crash prediction and risk management.

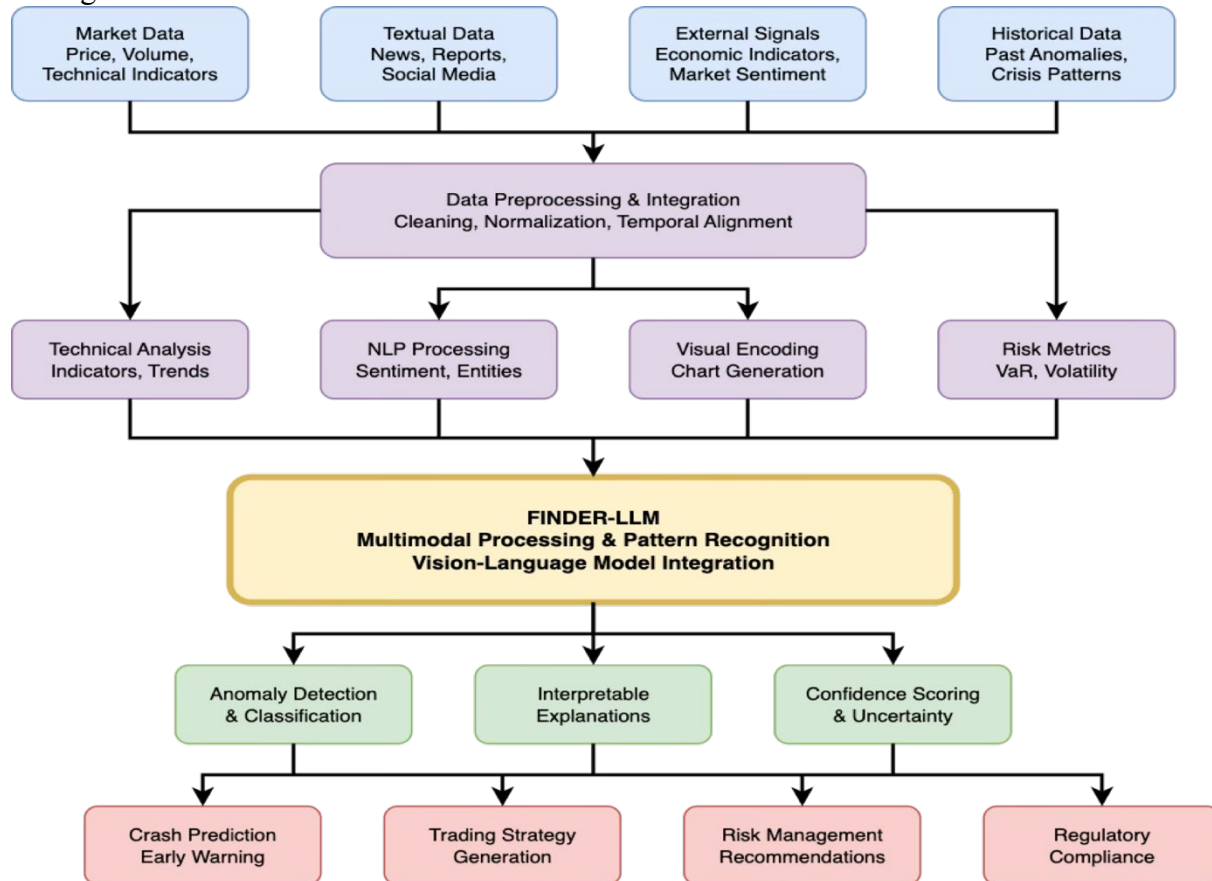


Figure 1. Pipeline of the FINDER-LLM Multimodal Anomaly Detection Framework

Source: Authors' own research.

Data Processing and Rolling Window Segmentation

The foundation of our analysis is daily market data, comprising Open, High, Low, Close, and Volume vectors. Let $\mathbf{v}_t = [OHLCV_t]^\top$ denote the vector of daily prices at time t , for $t = 1, \dots, T$. To capture the temporal evolution of market regimes without look-ahead bias, we employ a rolling window approach. We define a window length w . For each time step i ranging from w to T , the system constructs a rolling window segment W_i :

$$W_i = \{\mathbf{v}_{i-w+1}, \dots, \mathbf{v}_i\} \quad (1)$$

the segment W_i serves as the base unit for all subsequent feature engineering and visualization steps.

For each window W_i , we construct a high-dimensional feature set that captures trend, volatility, and risk metrics. These features are generated both as numerical vectors for statistical validation and as visual elements for the LLM.

Technical Indicators

We compute a vector of technical indicators $X_i \in \mathbb{R}^d$ to quantify market behavior. Based on standard financial literature and practical trading setups, we select indicators that cover diverse aspects of market dynamics.

Let the closing price at time t be defined as P_t , and the corresponding trading volume is represented as V_t . The logarithmic return is denoted as r_t . The historical volatility (HV_w) calculates the annualized standard deviation of logarithmic returns over specified window $w \in \{10, 30, 120\}$

$$HV_w = \sqrt{252} \text{std}(r_{t-w+1}, r_{t-w+2}, \dots, r_t) \quad (2)$$

The Average True Range (ATR) over a 14-day period measures market volatility by applying an exponentially weighted moving average (EWMA) to true range (TR_t)

$$ATR = EWMA_{14}(TR_t), \quad TR_t = \max\{H_t - L_t, |H_t - P_{t-1}|, |L_t - P_{t-1}|\} \quad (3)$$

The Moving Average Convergence Divergence (MACD) identifies momentum by calculating the difference between the 12-period and 26-period exponential moving averages (EMA) of the closing price.

$$MACD = EMA_{12}(P_t) - EMA_{26}(P_t) \quad (4)$$

The MACD Histogram represents the divergence between the MACD itself and its 9-period EMA signal line.

$$MACD \text{ Histogram} = MACD_t - EMA_9(MACD_t) \quad (5)$$

The Relative Strength Index (RSI) over a 14-period window evaluate price momentum using the ratio of average gains to average losses.

$$RSI = 100 - \frac{100}{1 + RS_t}, \quad RS_t = \frac{AvgGain_{14,t}}{AvgLoss_{14,t} + \varepsilon} \quad (6)$$

These average gains and losses are computed using a 14-period EWMA of the positive and negative price changes.

$$AvgGain, AvgLoss = EWMA_{14} \text{ of } \max(P_t - P_{t-1}, 0), \max(P_{t-1} - P_t, 0) \quad (7)$$

Bolinger Bands (BB) establish upper and lower price channels based on a 20-period simple moving average (SMA) and a two-standard-deviation threshold.

$$BB_{upper}, BB_{lower} = SMA_{20}(P_t) \pm 2 \text{std}_{20}(P_t), \quad SMA_k(P_t) = \sum_{i=0}^{k-1} P_{t-i} \quad (8)$$

The Bolinger Bands Width normalizes the distance between these upper and lower bands relative to the central SMA.

$$BB \text{ Width} = \frac{BB_{upper} - BB_{lower}}{SMA_{20}(P_t) + \varepsilon} \quad (9)$$

Death and Golden Crosses provide structural trend signals when the 50-period SMA intersects the 200-period SMA from above or below. The Maximum Drawdown (MaxDD) over a 20-day window captures the largest percentage drop from a rolling peak.

$$MaxDD_{20d} = \min_{k=0,\dots,19} \left(\frac{P_{t-k}}{\max(P_{t-k-19}, \dots, P_{t-k})} - 1 \right) \quad (10)$$

The 14-day Ulcer Index measures downside risk by aggregating the squared percentage retracements from recent highs.

$$Ulcer\ Index_{14d} = \sqrt{\frac{1}{14} \sum_{j=0}^{13} \left(\frac{P_{t-j} - \max(P_{t-j-13}, \dots, P_{t-j})}{\max(P_{t-j-13}, \dots, P_{t-j}) + \varepsilon} \times 100 \right)^2} \quad (11)$$

PICBE |
196

Value at Risk (VaR) estimates the return quantile at specific probability levels $p \in \{5\%, 1\%\}$.

$$VaR_p = q_p(r_t) \quad (12)$$

Expected Shortfall (ES) calculates the conditional expectation of returns that fall below the 5% VaR threshold

$$ES_{0.05} = E[r_j \mid r_j \leq VaR_{0.05}] \quad (13)$$

On-Balance Volume (OBV) accumulates trading volume based on the direction of daily price changes to assess buying and selling pressure.

$$OBV_t = OBV_{t-1} + \begin{cases} V_t, & P_t > P_{t-1} \\ -V_t, & P_t < P_{t-1} \\ 0, & P_t = P_{t-1} \end{cases} \quad (14)$$

These indicators are not only fed into the model as numerical summaries but are also rendered into a multi-panel chart image I_i . This visualization aligns the indicators vertically, allowing the Vision-Language Model to visually inspect correlations (e.g., a price drop coinciding with a spike in volatility) in a manner similar to human technical analysts.

Textual and News Integration

To provide the necessary narrative context, we integrate unstructured textual data. For each day in the window, we retrieve relevant financial news headlines and reports. These are aggregated into a text object N_i . This textual component is crucial for differentiating between statistical anomalies (noise) and event-driven market shifts (signal), such as those caused by geopolitical events or macroeconomic announcements.

Prompt Construction and Inference Strategy

The core innovation of FINDER-LLM lies in its prompting strategy. We construct a multimodal prompt $\wp_i$ that combines the visual, tabular, and textual modalities. We investigate the zero-shot prompting paradigms, the model is presented with the current window's data tuple $\wp_i^{zero} = (I_i, S_i, N_i, C_i)$, where I_i is the chart image, S_i is the tabular summary of indicators, N_i is the news context, and C_i represents the historical price context. The model must rely solely on its pre-trained knowledge to identify anomalies. For complete transparency and to ensure reproducibility, the full system prompt structure and a representative pseudocode example are publicly accessible via our Quantlet repository .

The inference process, denoted as $\mathcal{F}_{LLM}(\wp_i)$, yields a structured output containing the predicted anomaly label $\hat{y}_i$, the probability score of the anomaly $\hat{p}_i$, the classification of the anomaly type $\hat{c}_i$, and a natural language rationale $\hat{r}_i$ explaining the decision.

LLM Embedding Extraction: For each prompt φ_i , after generating the LLM response, we extract a latent representation that encodes the model's holistic reasoning over the multimodal input. Let the embedding function be denoted as $\mathcal{E}(\cdot)$, which maps the prompt and response pair to a continuous vector. The embedding vector $\mathbf{z}_i$ is obtained via a dedicated embedding model applied to the serialized LLM output.

Ground Truth Definition and Labeling

To evaluate the performance of the model, we establish a strictly quantitative definition for market crashes. We define the target variable y_i based on the future log return over a specific horizon h . Let $R_{i,h}^{fwd,log}$ be the future log return:

$$R_{i,h}^{fwd,log} = \log\left(\frac{C_{i+h}}{C_i}\right) \quad (15)$$

An observation is labeled as a crash ($y_i = 1$) if the future return falls below a dynamic statistical threshold. Specifically, we use the rolling quantile of the previous w log returns:

$$y_i = \begin{cases} 1, & \text{if } R_{i,h}^{fwd,log} < q_\alpha \\ 0, & \text{otherwise} \end{cases} \quad (16)$$

where q_α represents the α quantile of the historical distribution. In practice, to provide sufficient historical context while remaining responsive to recent market conditions, we set the rolling window size to 126 trading days, representing approximately six months of activity. Furthermore, we evaluate the detection capability under two distinct levels of market stress by setting the quantile parameter to 0.01 for extreme crashes and 0.05 for moderate crashes. This dynamic labeling ensures that the system adapts to changing volatility regimes, flagging only the most extreme negative deviations as true anomalies.

4 Empirical Results

To validate the efficacy of the FINDER-LLM framework, we conducted a comprehensive proof-of-concept study using the S&P 500 index, a widely recognized benchmark for the US equity market. Our dataset spans a twenty-year period from January 1, 2005, to June 30, 2025. This extensive timeframe allows us to cover multiple market regimes, including the 2008 Global Financial Crisis, the 2020 COVID-19 crash, and the subsequent recovery phases. We adopted a strict temporal splitting strategy to prevent data leakage and ensure a realistic evaluation of the model's predictive power. The data is divided into two distinct sets. The Training and Calibration Set from January 1, 2005, to June 30, 2024, aligns with the knowledge cutoff of the GPT-4.1 model. We use this historical data to calibrate the probability thresholds to calibrate probability thresholds and to refine the prompting setup used in the study. The Out-of-Sample Test Set, July 1, 2024, to June 30, 2025, represents strictly unseen data for the model. All performance metrics reported in this section are derived exclusively from this test set to assess the model's ability to generalize to future market conditions.

The rolling window size was set to $w = 126$ trading days (approximately six months), providing sufficient historical context for the technical indicators to stabilize. We evaluated the detection capability against two levels of market stress: a moderate crash threshold defined by the 5% quantile of historical log-returns, and an extreme crash threshold defined by the 1% quantile.

Qualitative Case Study: The Lehman Brothers Crisis

Before analyzing the aggregate statistical performance, it is instructive to examine the model's reasoning process in a specific historical context. Figure 2 presents the generated technical anomaly chart and the FINDER-LLM output for the S&P 500 on September 12, 2008, the trading day immediately preceding the collapse of Lehman Brothers.

In this instance, the FINDER-LLM framework output an Anomaly Score of 0.85 and correctly assigned an Anomaly Label of 1. Crucially, the model provided a structured rationale that mirrors the deductive process of a human analyst.

Based on these signals, the model concluded that the market was undergoing a "regime shift or pre-crash environment," accurately predicting the high-risk instability that characterized the subsequent trading days. This qualitative evidence demonstrates the model's capacity to synthesize disparate technical signals into a coherent risk narrative.

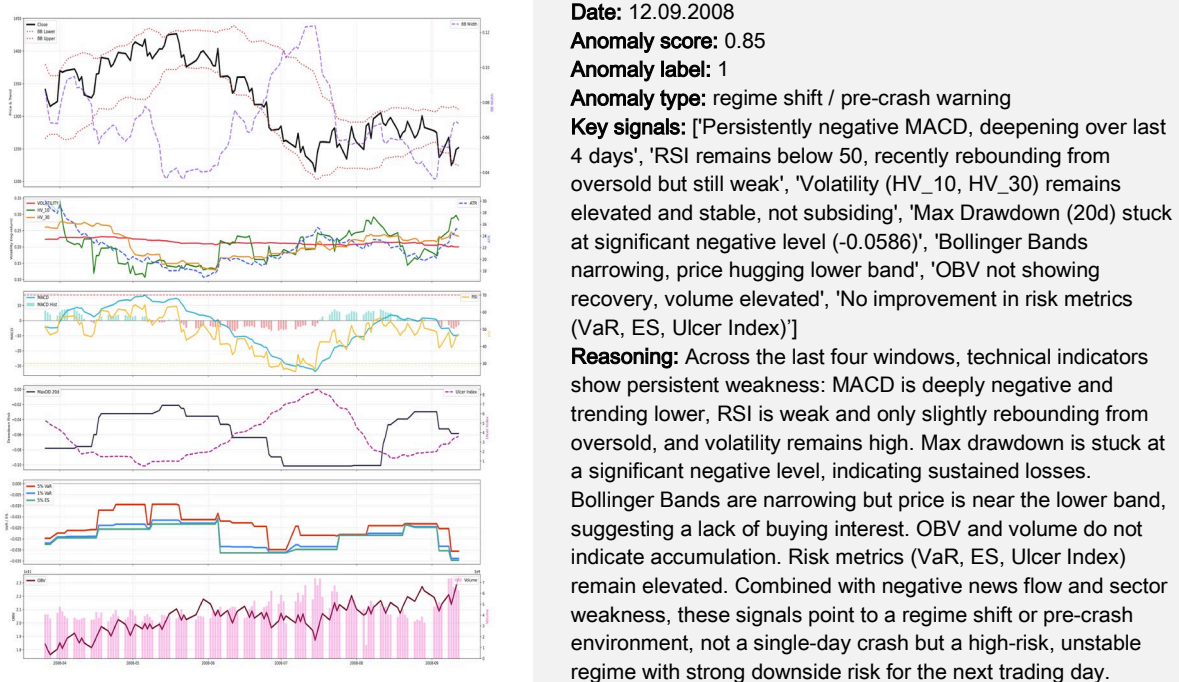


Figure 2. FINDER-LLM Output on the S&P 500 During the 2008 Pre-Crisis Period

Source: Authors' own research.

Quantitative Performance Analysis

We benchmarked the FINDER-LLM against a machine learning baseline, specifically a Long Short-Term Memory (LSTM) model processing the engineered technical indicator vector. The evaluation metrics focused on the ability to classify days into normal or anomaly categories. We report the Receiver Operating Characteristic Area Under the Curve (ROC AUC) and Accuracy to evaluate the overall classification capabilities. To provide a more robust evaluation, we include a confusion matrix analysis, detailing True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN), alongside precision and recall to evaluate the overall classification capabilities. Table 1 summarizes the performance on the out-of-sample test set for both the 1% and 5% quantile crash definitions.

Table 1. Out-of-Sample Classification Performance under 1% and 5% Crash Quantile

Approach	ROC AUC	Accuracy	TP	FP	TN	FN
$\alpha = 0.01$						
LSTM	0.31	0.20	4	105	142	16
LLM Zero-Shot (Embedding)	0.66	0.85	9	154	93	11
LLM Zero-Shot (Prompt)	0.59	0.96	1	33	225	6
$\alpha = 0.05$						
LSTM	0.42	0.14	9	154	93	11
LLM Zero-Shot (Embedding)	0.42	0.38	0	4	236	7
LLM Zero-Shot (Prompt)	0.58	0.83	4	26	201	16

Source: Authors' own research.

The LSTM baseline performed poorly, yielding notably low ROC AUC and accuracy metrics across both thresholds. This underperformance highlights the limitations of relying strictly on numerical patterns without contextual understanding in complex market environments, the model often triggers signals on minor deviations that do not materialize into systemic crashes.

The LLM-based approaches demonstrated a significant improvement over the baseline. The Zero-Shot (Prompt) method, which asks the model to output a score and label directly, achieved the highest overall accuracy, particularly at the stricter 1% threshold. However, the confusion matrix shows that this high accuracy is partly due to the imbalanced data, the model correctly identifying the majority class (normal days), as it produced very few false positives but struggled to capture true positives at this extreme level. At the 5% threshold, it successfully identified multiple true positives while keeping false positives relatively low, maintaining a balanced ROC AUC and suggesting a stable discriminative ability.

Interestingly, the Zero-Shot (Embedding) approach achieved the highest ROC AUC for the extreme 1% threshold. Although it generated more false positives than the prompt method, its confusion matrix demonstrates a better capacity to capture risk signals. This suggests that the latent representation of the model's reasoning contains richer information than the explicit probability score alone. However, its performance degraded significantly at the 5% threshold, returning a confusion matrix largely similar to the baseline. This indicates that embeddings might be more sensitive to extreme tail events than to moderate volatility.

5 Limitation and Discussion

While the FINDER-LLM framework represents a step forward in interpretable financial AI, several limitations must be acknowledged. First, there is an inherent dependence on the specific reasoning capabilities of the underlying LLM. We observed that anomaly scores and textual explanations can shift due to minor perturbations in the prompt phrasing, raising concerns about deterministic reliability.

Second, the Training Data Cutoff presents a challenge. The model's internal knowledge base ends in June 2024. While our pipeline feeds recent news via the prompt, the model lacks intrinsic knowledge of global events occurring after the cutoff, potentially limiting its contextual understanding of novel geopolitical crises.

Third, the system does not currently process raw fundamental data, for example, balance sheets, earnings call transcripts, or real-time macro-economic data streams. It relies solely on the technical features and news headlines provided in the prompt. This abstraction may miss deeper structural risks buried in financial statements.

Finally, the computational cost of generating daily charts and querying a high-intelligence vision model at scale is non-trivial compared to standard statistical methods. This currently restricts the framework's applicability to high-frequency trading scenarios.

6 Conclusion

This study introduces FINDER-LLM, a novel multimodal framework designed to improve financial anomaly detection by treating market regime classification as a visual and textual recognition task. By integrating aligned price data, technical charts, and financial news, this approach moves beyond previous single-modality models and establishes a practical evaluation structure using real-world market data. The results demonstrate that LLM-based anomaly detection provides interpretable daily risk signals with moderate predictive accuracy, offering the structured, structured and transparent explanations that traditional models often lack. Specifically, zero-shot prompt models outperform classical technical indicators, which show limited predictive power in this complex market setting. Furthermore, embedding-based approaches improve the robustness of the framework, and the pipeline is designed to be modular and accessible to support further development. However, overall anomaly detection remains challenging, indicating that further research is needed to enhance LLM performance for reliable financial risk forecasting.

While the FINDER-LLM framework presents a novel approach, practical challenges remain beyond the data and computational constraints discussed previously. Most notably, the system's ability to handle missing or highly noisy data requires further investigation. To address these challenges, further work will focus on several key directions. First, we plan to expand the pipeline to multi-asset and global markets, applying the methodology to equities, bonds, and cryptocurrencies worldwide. Second, refining probability calibration and model tuning is essential to improve thresholding and reduce false positives. Additionally, conducting an ablation study to identify which technical indicators are most essential for the LLM will be crucial. Finally, future efforts will aim to make LLM explanations clearer and more auditable to boost interpretability and trust, while also optimizing the system for real-time, large-scale deployment to lower operational costs and streamline daily operations.

Acknowledgements

This paper is supported through the project "IDA Institute of Digital Assets", CF166/15.11.2022, contract number CN760046/ 23.05.2023; the project "AI for Energy Finance (AI4EFin)", CF162/15.11.2022, contract number CN760048/23.05.2023, financed under the Romania's National Recovery and Resilience Plan, Apel nr. PNRR-III-C9-2022-I8; and the Marie Skłodowska-Curie Actions under the European Union's Horizon Europe research and innovation program for the Industrial Doctoral Network on Digital Finance, acronym DIGITAL, Project No. 101119635.

This paper is co-financed by the Bucharest University of Economic Studies during the PhD program. During the preparation of this work, the author(s) utilized ChatGPT 5 to enhance readability and language clarity. Following the use of this tool, the author(s) meticulously reviewed and revised the content as necessary and assume(s) full responsibility for the final content of the publication.

References

- Chandola, Varun, Arindam Banerjee, and Vipin Kumar. 2009. "Anomaly Detection: A Survey." *ACM Computing Surveys* 41 (3): 1–58. <https://doi.org/10.1145/1541880.1541882>
- Bauwens, Luc, Sébastien Laurent, and Jeroen V. K. Rombouts. 2006. "Multivariate GARCH Models: A Survey." *Journal of Applied Econometrics* 21 (1): 79–109. <https://doi.org/10.1002/jae.842>
- Wipplinger, Evert. 2007. "Philippe Jorion: Value at Risk—The New Benchmark for Managing Financial Risk." *Financial Markets and Portfolio Management* 21 (3): 397. <https://doi.org/10.1007/s11408-007-0057-3>
- Cortes, Corinna, and Vladimir Vapnik. 1995. "Support-Vector Networks." *Machine Learning* 20 (3): 273–297. <https://doi.org/10.1007/BF00994018>
- Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32. <https://doi.org/10.1023/A:1010933404324>
- Zhou, Zihao, and Rose Yu. 2024. "Can LLMs Understand Time Series Anomalies?" *arXiv preprint arXiv:2410.05440*. <https://doi.org/10.48550/arXiv.2410.05440>
- Xu, Xiong Xiao, Haoran Wang, Yueqing Liang, Philip S. Yu, Yue Zhao, and Kai Shu. 2025. "Can Multimodal LLMs Perform Time Series Anomaly Detection?" *arXiv preprint arXiv:2502.17812*. <https://doi.org/10.48550/arXiv.2502.17812>
- Park, Taejin. 2024. "Enhancing Anomaly Detection in Financial Markets with an LLM-Based Multi-Agent Framework." *arXiv Preprint arXiv:2403.19735*. <https://doi.org/10.48550/arXiv.2403.19735>
- Yang, Alan, Yulin Chen, Sean Lee, and Venus Montes. 2025. "Refining Time Series Anomaly Detectors Using Large Language Models." *arXiv preprint arXiv:2503.21833*. <https://doi.org/10.48550/arXiv.2503.21833>
- Tsai, Che-Ping, Ganyu Teng, Phillip Wallis, and Wei Ding. 2025. "AnoLLM: Large Language Models for Tabular Anomaly Detection." In *The Thirteenth International Conference on Learning Representations*.
- Liu, Haoxin, Shangqing Xu, Zhiyuan Zhao, Ling kai Kong, Harshavardhan Prabhakar Kamarthi, Aditya Sasanur, Megha Sharma, Jiaming Cui, Qingsong Wen, Chao Zhang, and others. 2024. "Time-MMD: Multi-Domain Multimodal Dataset for Time Series Analysis." *Advances in Neural Information Processing Systems* 37: 77888–77933. <https://doi.org/10.52202/079017-2476>
- Wang, Ziao, Yuhang Li, Junda Wu, Jaehyeon Soon, and Xiaofeng Zhang. 2023. "FinVis-GPT: A Multimodal Large Language Model for Financial Chart Analysis." *arXiv preprint arXiv:2308.01430*. <https://doi.org/10.48550/arXiv.2308.01430>
- Zhang, Yang, Wenbo Yang, Jun Wang, Qiang Ma, and Jie Xiong. 2025. "CAMEF: Causal-augmented Multi-modality Event-driven Financial Forecasting by Integrating Time Series Patterns and Salient Macroeconomic Announcements." In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 3867–3878. <https://doi.org/10.1145/3711896.3736872>