**Reviewer #1: The only element of the article I like is the idea to use LLM to process news articles, as you describe:**  
 *"Each article i ∈ D implies a conversation with the LLM. The structure of the conversation implies defining first a "system message", which provides a general context and purpose to the model. In our case:  
− You are a function calling LLM that analyses business news in Spanish.  
− For every article, you must identify the firms directly affected by the news. Do not include every firm mentioned in the article, only include those that are directly affected by the shocks narrated therein.  
− The identified firms must be Spanish and should be publicly listed in the Spanish exchange (their ticker is of the form 'TICKER.MC'). Do not include non-Spanish foreign firms. Do not include Spanish firms that are not publicly traded.  
− For each identified firm, classify the shocks that affect them (type, magnitude, category). The type of shock can be 'demand', 'supply', 'financial', 'policy', or 'technology'. The magnitude can be 'minor' or 'major'. The direction can be 'positive' or 'negative'.  
− If a firm is affected neutrally by the news article, don't include it in the analysis."*  
  
**This idea is meaningful, and the paper could be good if the rest of the manuscript is proper. However, even though I can see that there is a lot of work behind this manuscript, most of the other elements of the manuscript are weak/strange/inappropriate.**  
  
**To be more specific:**Abstract and conclusion

**Let's start with the abstract. It is too general. After reading the abstract, I learned too little about the article, the findings, etc. Your main conclusion seems to be:**  
  
*"Using a simple trading strategy, we demonstrate that our LLM-based classification significantly outperforms a benchmark based on clustering vector embeddings, generating consistent profits out-of-sample while maintaining transparent and durable trading signals."*  
  
**The conclusion is better. There I can understand more of what you do. Here, your overall conclusion seems to be:**  
 *"Moreover, the resulting trading signals are both long-lasting and economically relevant, as they are based on fundamental economic shocks rather than statistical patterns. The results show that the LLM-based trading strategy effectively identifies winners and losers, illustrating the parser's ability to anticipate market trends by comprehending the economic implications of firm-specific shocks. This approach generates a consistent profile of earnings in the test set, with results robust to the choice of hyperparameters-the holding period length of the trading strategy and the number of selected clusters for trading. Our findings demonstrate a promising avenue: LLMs, when guided by appropriate economic frameworks, can help predict market reactions to news through systematic classification of economic shocks embedded in financial narratives."*

Overselling your paper  
**Altogether, after reading your abstract and conclusion, I got the impression that you managed to apply LLM to news articles in such a way that you designed some very good, profitable, and stable trading strategy. It is just surprising that the economic magnitude of this great trading strategy is not mentioned in either the abstract or conclusion.  
Later, after reading the whole article, I find in the results that:**  
  
*"Specifically, while KMeans strategies generate significant losses out-of-sample (-20.0% to -23.6% annually), our LLM-based approach achieves near-neutral to slightly positive net returns (-1.5% to +3.1%) in the test period,"*

**This means that your trading strategy practically does not work. In your own words, it delivers near-neutral to slightly positive net returns.**

Implementation concerns  
  
**Could your strategy be implemented in real time?**  
**Moreover, your paper does not discuss properly how computationally demanding are your calculations. Could you implement your strategy in a real time? It seems not. In the data section, you write:**  
*"This paper employs a dataset of Spanish business news articles sourced from Dow Jones Newswires, covering the period from June 24, 2020, to September 30, 2021. The selection of this timeframe is deliberate, driven by two key considerations. First, given the substantial computational demands of LLM-based analysis, we strategically focus on a smaller, carefully curated dataset. This deliberate scope reduction allows us to thoroughly demonstrate our novel methodology's effectiveness in decoding market-news relationships while keeping computational costs manageable."*

**In other words, you deliberately did not choose the largest dataset possible but instead a smaller dataset, just a little bit more than one year, to keep computational costs manageable. My intuition tells me that since you needed to do this, you could not implement your strategy in real-time. However, this should be more discussed in your paper.**  
  
  
False claim about generalizability of your trading strategy  
**Following text in the data section is also problematic:**  
  
*"Second, we specifically chose the Covid-19 era to test our methodology's extrapolative capabilities during periods of significant market instability and volatility. While existing textual algorithms typically perform well in stable market conditions, they often struggle to generalize effectively during periods of heightened uncertainty. By focusing on this volatile period, we can better assess our methodology's robustness and its ability to maintain predictive power under challenging market conditions."* **When you make statement like "While existing textual algorithms typically perform well in stable market conditions, they often struggle to generalize effectively during periods of heightened uncertainty. ", you really need to cite relevant literature.**  
  
**As far as I know the literature, it is easier to find patterns (and therefore trading strategies) in the stock market during the crisis period, see e.g. articles:  
Sentana, E., & Wadhwani, S. (1992). Feedback traders and stock return autocorrelations: evidence from a century of daily data. The Economic Journal, 102(411), 415-425.  
Kim, J. H., Shamsuddin, A., & Lim, K. P. (2011). Stock return predictability and the adaptive markets hypothesis: Evidence from century-long US data. Journal of Empirical Finance, 18(5), 868-879.  
Lyócsa, Š., & Molnár, P. (2020). Stock market oscillations during the corona crash: The role of fear and uncertainty. Finance Research Letters, 36, 101707.  
  
Therefore, even if you find some predictability during the COVID-19 era, it does not mean that your results generalize to other time periods. In other words, your statements like "By focusing on this volatile period, we can better assess our methodology's robustness" are false.**  
  
Bad benchmarking  
**In your paper, you compare LLM-based approach with K-means clustering. Why should the approach based on K-means clustering be a benchmark? Is it some commonly used approach? I do not think so. Is there a reason we should expect this approach to work well? I do not think so, and your results confirm that this approach actually performs very poorly. Therefore, keeping it as a benchmark is a very strange approach.  
Moreover, in several places in your manuscript, you criticize sentiment analysis in several places in your manuscript:**"First, it offers greater granularity and sophistication compared to traditional methods like sentiment analysis and topic modeling."  
"Sentiment analysis, while straightforward, lacks the necessary granularity, offering only positive, negative, or neutral classifications, which is insufficient to compare with our granular LLMbased economic shock classification. Additionally, sentiment analysis focuses on the emotional tone rather than the economic impact, it is prone to inconsistencies due to linguistic nuances and it can deliver very different outcomes depending on the specific sentiment analysis tool employed."  
  
**As far as I know, sentiment analysis has been frequently used in finance. If you argue that your LLM-based approach is good, it would be meaningful to compare it to a trading strategy based on sentiment analysis.**

Sharpe Ratio misuse  
**The second methodological issue is the way you use the Sharpe ratio. Sharpe ratio is generally a meaningful measure of the performance of a portfolio, not a single stock. Calculating Sharpe ratio for a single stock is not common, and does not make much sense. It makes even much less sense to calculate Sharpe ratio from 4 observations, as you do. You literally calculate Sharpe ratio for 4 days following the news article.  
Even more importantly,** **it would be shameful even for a bachelor student to calculate Sharpe ratio from abnormal returns. Did not know that you can calculate Sharpe ratio of the whole stock market? Following your calculations, it would not be possible. Abnormal return of the stock market is by definition always zero. Therefore, mean return would be zero, standard deviation would be zero, and Sharpe ratio would not be defined. This inconsistency arises because you do not know even such a basic definition what a Sharpe ratio is.**  
Lack of clarity **Altogether, the paper is very difficult to follow. Authors use mathematical notation, and particularly sets, way too much, even there where it is not necessary. This makes article very difficult to understand. For example, I do not understand how you calculate beta. Specifically, over which observations you estimate the CAPM equation. You say that you estimate it "over some window of time", but it is really not sufficient. You need to say over which window of time. There is some chance that this information is provided somewhere else, but this is really not a way how to write the article.**