Evaluating Stock Recommendations from Generative AI: An Experiment Using Deepseek for Generation and Grok for Auditing

Trinh Tuan Ngoc Bao, Nguyen Vu Quang Anh, Phan Tran Manh Cuong
VNU University of Engineering and Technology, Hanoi, Vietnam
ngocbaotrinhtuan@gmail.com

Abstract

Large Language Models (LLMs) are pioneering new methodologies in investment advisory. This study explores a fully automated pipeline where we utilize one generative AI (Deepseek) to produce stock recommendations from financial news, and subsequently use another AI (Grok) to evaluate the quality of those recommendations. Our objective is to test the feasibility of using AI not only for idea generation but also for their screening and validation, simulating an analysis workflow without human financial experts. The results show that Deepseek is capable of complex contextual reasoning and identifying potential investment themes. However, its recommendations still contain factual errors and hallucinations. Notably, the Grok model acted effectively as an "AI auditor," successfully identifying most of the weaknesses in Deepseek's output. This work demonstrates the potential of an "AI-audits-AI" model as an initial screening step but also underscores that human oversight remains indispensable for final investment decisions.

1 Introduction

The evolution of generative AI has spurred a wave of innovation in the financial industry. Previous works, such as that of Fang et al. [2023], have demonstrated that LLMs can extract predictive signals from financial news. This raises a subsequent question: Can we automate the entire cycle from reading news and generating investment ideas to preliminarily validating those ideas?

This study addresses that question by proposing and testing a two-step process. First, we assign Deepseek the role of an "Idea-Generating AI," tasking it with reading news and proposing specific stocks. Second, instead of relying on human financial experts for evaluation, we deploy an "Auditor AI," Grok, with the mission of objectively validating Deepseek's recommendations.

Our main contribution is the presentation of a novel methodology: using one AI to evaluate another in a complex financial task. We not only assess Deepseek's capabilities but also test whether Grok can serve as a reliable quality gatekeeper.

2 Related Work

The foundation of this research lies in two areas. The first is the application of NLP in finance, from dictionary-based methods [Loughran and McDonald, 2011] to transformer models like BERT and more recent work on ChatGPT in sentiment analysis [Fang et al., 2023]. The second is the field of research on the reasoning and fact-checking capabilities of LLMs themselves, where models are used

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

to self-correct or evaluate the outputs of other models. Our research is situated at the intersection of these two fields, applying the "AI-audits-AI" model to the practical problem of investment advisory.

3 Methodology

Our experimental process consists of two main stages: (1) The Recommendation Generation Stage by Deepseek and (2) The Quality Auditing Stage by Grok.

3.1 Stage 1: Recommendation Generation by Deepseek

We utilized the `deepseek-r1:7b` model accessed via a local server. To trigger the model's analytical capabilities, we designed a detailed role-playing prompt, instructing it to act as an expert.

> [title=Prompt sent to Deepseek] Assume you are a senior financial analyst. Please carefully read the following news from the Wall Street Journal. Based on the content of the news, if you would advise buying any stocks, write 'YES' and list 5 NYSE or Nasdaq stock names along with their ticker symbols. If your answer is simply 'NO', please briefly explain why.

3.2 Stage 2: Recommendation Auditing by Grok

After Deepseek generated a list of recommendations, we forwarded the entire output (including the original article and Deepseek's recommendation) to Grok for evaluation. The goal was for Grok to act as a second, critical-thinking analyst.

> [title=Prompt sent to Grok] You are a risk management and fact-checking expert in the financial sector. Below is a news article and an investment recommendation generated by another analytical AI. Your task is to rigorously evaluate this recommendation based on three criteria:
>
> 1. **Relevance:** Are the recommended stocks truly relevant to the article's content?
> 2. **Reasonableness:** Is the (even implicit) rationale behind selecting these stocks logical and economically sound?
> 3. **Factual Accuracy:** Are the company names and ticker symbols correct? Are there any "hallucinated" companies?
>
> Provide your analysis and conclude whether this recommendation is "Reliable" or "Requires Review".
>
> ---
>
> *[Original News Article Text]*
> *[Recommendation from Deepseek]*

4 Experiments and Results

We applied the above process to a set of financial news articles. A preliminary backtest on Deepseek's recommendations showed positive alpha signals, indicating the predictive potential of the method. However, the focus of this study is the qualitative analysis through Grok's auditing.

4.1 Case Study: Cross-Evaluation between Deepseek and Grok

We analyze a typical example concerning a news article about U.S. solar tariff policies.

- **Deepseek's Output:** Deepseek responded with 'YES' and proposed 5 stocks, including First Solar (FSLR), NextEra Energy (NEE), Siemens Solar Cells (SNEURY), and Perion Power (PRW).
- **Grok's Analysis:**

  - *On Relevance and Reasonableness:* Grok agreed that recommending solar energy companies like First Solar and NextEra Energy was highly relevant and directly related to the news. Grok acknowledged Deepseek's good contextual reasoning.
  - *On Factual Accuracy:* This is where Grok demonstrated its "auditor" role. It pointed out critical errors:
    1. **Repetition:** Recommending "First Solar" twice.
    2. **Outdated/Incorrect Information:** Grok noted that "Siemens Solar Cells" is no longer an independent entity and the ticker SNEURY belongs to Siemens Energy, a less direct play.
    3. **Hallucination:** Grok identified that "Perion Power (PRW)" is a non-existent company, and the ticker PRW belongs to an unrelated ad-tech company.
- **Grok's Conclusion:** Grok concluded that the recommendation "Requires Review" due to serious factual errors, even though the initial analytical direction was correct.

This result shows that the two-AI model worked as expected: Deepseek generated a promising but raw idea, and Grok effectively screened and detected the flaws.

5 Limitations and Future Work

The biggest limitation of the "AI-audits-AI" approach is the risk of **correlated bias**. Both Deepseek and Grok are trained on large datasets and may share the same "blind spots" or biased tendencies. A mistake that both AIs fail to recognize would go undetected in this process.

Furthermore, Grok's evaluation is not yet a "gold standard." It can still make mistakes.

Future work should include:

1. **Human-in-the-Loop Validation:** Compare Grok's evaluation results with those of human financial experts to verify the reliability of the "AI auditor."
2. **Database Integration:** Integrate the LLM with structured financial databases (like Bloomberg, Refinitiv) for automated fact-checking instead of relying on another LLM.

6 Conclusion

Our research demonstrates the feasibility of an automated pipeline using one AI (Deepseek) to generate investment recommendations and another AI (Grok) to audit them. Deepseek showed its ability to identify investment opportunities from news, while Grok acted as an effective first-layer filter to detect basic errors.

This "AI-audits-AI" model opens a promising path to accelerate the process of generating and screening investment ideas. However, it cannot completely replace the deep judgment and due diligence of humans. The most appropriate role for this process today is as a powerful support tool, helping analysts filter the most promising ideas from a vast amount of information before they apply their expertise to make the final decision.

References

Hui Fang, Zhou Lu, and Sixuan Li. ChatGPT, Generative AI, LLMs, and Investment Advisory. *Available at SSRN 4480662*, 2023.

Tim Loughran and Bill McDonald. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65, 2011.