3.2 Textual Data

Textual data such as news or financial reports are critical to financial traders. We found all LLM powered agents reviewed in this paper use textual financial data as input. Based on the terminology commonly used in the financial industry, we categorize textual data into two types: Fundamental Data and Alternative Data.

3.2.1 Fundamental Data. Fundamental data encompasses information that represents the primary characteristics and financial metrics for assessing the stability and health of an asset. Fundamental data used in LLM trading agents includes financial reports and analyst reports.

Financial Reports. Financial reports, such as Form 10-Q and Form 10-K filings, are critical for understanding a company's performance. These documents provide LLM agents with insights into corporate financial status, performance, and future expectations. They are extensively utilized by financial trading agents like FinMem [53], TradingGPT [27], and FinAgent [57]. These works incorporate financial reports to enrich the agents' memory and make informed trading decisions.

Analyst Reports. In addition to financial statements, analyst reports and investment research from industry professionals provide invaluable data. These sources offer high-quality insights, opinions, and forecasts beyond the information found in public financial reports and news articles. For example, FinAgent [57] incorporates expert guidance from SeekingAlpha[1] as a crucial input to its decision-making module. It is used with other data sources such as market intelligence, analysis of price movements, and historical trading decisions.

3.2.2 Alternative Data. Alternative data refers to non-traditional information used to evaluate companies and markets. This type of data complements traditional sources such as financial reports. By leveraging alternative data, trading agents can obtain unique perspectives on various issues, thereby enhancing their investment decision-making processes.

News Data. News data from reputable sources such as Bloomberg[2], The Wall Street Journal[3], CNBC Television[4], or stock research platforms provides real-time information on market movements, industry trends, and company-specific developments. This type of data is extensively used in various studies [20, 27, 51, 53, 57] to stay up-to-date with the real-world financial market. Specifically, LLMs excel at extracting sentiment information from news data, which could a crucial signal for trading decisions.

Social Media Data. In addition to traditional news sources, researchers can also use of social media data, such as Twitter, StackExchange, StockTwits and Reddit posts, to capture more informal, real-time discussions about financial topics. There are many machine learning models that ultilize social media data for stock price prediction such as [14, 40, 58]. However, SEP [19] is the only work we've reviewed that incorporates real-time social media data in LLM trading agent. In SEP, LLM is used to generate and summarize key facts from twitter data of a given stock. Incorporating social media data is a under studied field but with great potential.

[1]https://seekingalpha.com/
[2]https://www.bloomberg.com/
[3]https://www.wsj.com/
[4]https://www.cnbc.com/

3.3 Visual Data

Numerical and textual data have been predominately used in trading agent design, while visual data has been less explored as an additional data source. One of the reason for this disparity stems from the challenges that existing LLM models have in effectively processing and understanding financial visual data. While recently proposed multimodal LLMs such as LLaVA [28], GPT-4v[35] possesses the capability to processing visual data, most of these models have not been specifically trained and evaluated on financial visual data such as Kline charts, volume charts. An early experiment by FinAgent [57] in incorporating visual data using GPT-4v in the trading agent context has shown promise. FinAgent integrates Kline Charts and Trading Charts along with numerical and textual data. It demonstrated significant trading performance improvements over FinMem which uses a similar architecture but without visual input. This effort represents a significant step forward in utilizing visual data within LLM frameworks for trading applications, by leveraging trading chart information, which forms the cornerstone of technical analysis widely used by traders. This pioneering work sets a promising direction for further exploration in integrating visual data into LLM-based trading agents.

3.4 Simulated Data

Simulated data and environments are created to replicate real-world scenarios, providing finance professionals with effective tools for understanding both market dynamics and LLM agent behavior. In [54], a group of LLM stock agents with varying personalities engage in trading within a simulated environment. This simulation includes not only market price fluctuations but also synthetic events such as interest rate changes and the release of financial reports. Additionally, agents can communicate via a Bulletin Board System. Experiments have shown that agents with different personalities react differently, and external factors significantly influence their behavior.

Simulated data is also invaluable for researching LLM behaviors concerning bias, ethics, and robustness under extreme circumstances in a controlled manner. [41] defined several real-world scenarios including one with extreme pressure to examine LLM agents' behavior in these circumstances. The study revealed that LLMs are capable of taking unethical actions under high-pressure conditions, such as using insider information to trade for profit and even crafting deceptive explanations to conceal such actions. This study underscores the potential regulatory risks associated with using LLMs in financial trading. Therefore, it is imperative to thoroughly investigate such issues before deploying them in a production environment.

4 EVALUATION

In the papers we have surveyed, LLM powered trading agent have demonstrated superior performance during backtesting. In this section, we discuss trading strategies generated by LLM agents, as well as the evaluation metrics and baselines used to assess the performance of LLMs via backtesting.

4.1 Trading Strategy

LLM generates simple trading signals such as "Buy", "Hold", "Sell" by analyzing textual data like market news or financial statements. In FinMem[53] and FinAgent[57], the signal is directly used for trading action for a particular stock. However, when managing a portfolio with multiple stocks, a common approach is to use ranking-based strategies. These strategies require a numeric score to rank the stocks and allocate funds based on the magnitude of these scores. In FinLlama[20], all stocks in S&P500 index are ranked by an LLM and top 35% are assigned to long position while the bottom 35% are assigned to the short position. A similar approach is adopted in [18, 29], where the long-short strategy has shown to outperform both long-only and short-only strategies in backtesting. On the other hand, [50] allocates long positions to stocks with overall positive news sentiment and short positions to those with negative sentiment, without considering the magnitude of the sentiment scores. This approach does not fully utilize the signal, leading to the observation that the long-short strategy performed worse than the long-only strategy in their experiment. In [56], stocks are grouped based on their signal rankings, with the top-ranked group showing the best returns compared to others.

In executing trading strategies, stocks are typically weighted either equally or based on market capitalization size. In both[18] and [29], portfolios weighted by market capitalization have shown slightly higher returns than those that are equally weighted. We conjecture that the quality of textual signals from large-cap companies is better than that from smaller companies due to the bias in news coverage.

4.2 Metrics

Portfolio Performance Metrics. Almost all works we surveyed uses common performance metrics in evaluating the trading agent. Cumulative return and annualized return are used to measure overall profitability of the trading strategies. Sharpe Ratio [44] and Maximum Drawdown are used to assess the risk of the trading performance. One observation we had is that while both risk and profit metrics are commonly used, few studies consider trading costs in their evaluations.

- Cumulative Return:

$$\text{Cumulative Return} = \left(\frac{P_t - P_0}{P_0}\right) \times 100\%$$

  where:
  - P_t is the ending price (or value) at time t
  - P_0 is the initial price (or value) at the beginning
- Annualized Return

$$\text{Annualized Return} = \left(\frac{P_t}{P_0}\right)^{\frac{1}{t}} - 1$$

  where:
  - P_t is the ending price (or value) at time t
  - P_0 is the initial price (or value) at the beginning
  - t is the number of years
- Sharpe Ratio:

$$\text{Sharpe Ratio} = \frac{R_p - R_f}{\sigma_p}$$

  where:
  - R_p is the return of the portfolio
  - R_f is the risk-free rate
  - σ_p is the standard deviation of the portfolio's excess return
- Maximum Drawdown up to time T is given by:

$$\text{MDD}(T) = \max_{\tau \in (0,T)} D(\tau) = \max_{\tau \in (0,T)} \left[\max_{t \in (0,\tau)} X(t) - X(\tau) \right]$$

  where:
  - $X(t)$ is the value of the portfolio at time t
  - $X(\tau)$ is the value of the portfolio at time τ
  - T is the time period being considered

Signal Metrics. Sometimes, portfolio performance metrics do not directly reflect the performance of a trading agent or the effectiveness of a trading signal. Therefore, it is equally important to monitor the predictive power of the generated signals. In [18] and [60], the F1 score and accuracy are used to measure the model's prediction accuracy of news sentiment. Meanwhile, [10] and [56] use the win rate to measure the proportion of profitable trades out of all executed trades. In QuantAgent [48], the Information Coefficient (IC) [55] is calculated to quantify the correlation between predicted signals and future returns.

System Metrics. Utilizing LLM-powered trading agents to process information and generate trading signals often involves leveraging commercial LLM APIs such as ChatGPT[5]. However, QuantAgent is the only study we have encountered that addresses the cost of generating LLM tokens and the computational time complexity for both training and inference. This could be due to the fact that the cost of token generation is usually negligible compared with the capital size of the portfolio.

4.3 Backtest Setting

Number of Years	Count of Papers
0∼2	8
2∼5	2
≥5	4

Table 1: Number of Years Covered in Backtest Testing

To evaluate the performance of LLM powered agents, most of the work use backtesting with real market data. For agents evaluated on single-stock portfolios, stocks with the highest volume of accessible news data are seleted for testing. For example, stocks such as TSLA, AMZN, MSFT, COIN, NFLX, GOOGL, META, PYPL selected to trade in [53], [57] and [60]. For agent managing multi-stock portfolios, index component stocks are typically selected, such as those from the SP500[45] and CSI300[5].

Most agent-based models are backtested exclusively on the stock market. Among the 14 papers that use real market data for backtesting, 9 focus on the US stock market and 5 on the Chinese market.

[5] https://chatgpt.com/

Only FinAgent [57] extends its backtesting on the cryptocurrency market, specifically trading ETH [4].

We also observe that most evaluations set the backtesting period between 2020 and 2024, coinciding with the publication date of the work. On average, the median of testing period is only 1.3 years (Table 1), with the exact start and end dates chosen rather arbitrarily. While LLM agents have demonstrated strong performance during backtesting, a short and single backtesting period may diminish the credibility of the results.

4.4 Baseline and Performance

During backtesting, the baselines methods can be divided into 3 major categories: Rule based, Machine Learning (or Deep Learning) based and Reinforcement Learning based. In [31, 53, 57], rule based strategy such as "Buy and Hold", "Mean Reversion" and "Short-Term Reversal" are used as baseline. Given that classification models can be used for news sentiment prediction, machine learning or deep learning models such as Random Forest[30], LightGBM[17], LSTM[13], and BERT[7] are also used as baselines. Furthermore, Reinforcement Learning algorithm are increasingly popular in quantitative trading [25]. Deep reinforcement learning frameworks such as PPO[43] and DQN[34] are also used as a benchmark in [53, 57].

Overall, LLM powered trading agents have demonstrated strong performance in backtesting. Our survey[8, 29, 53, 57] shows that LLM agents have achieved annualized return ranging from 15% to 30% over the strongest baseline during backtesting period with real market data, which demonstrates the great potential of using LLM in financial trading.

5 LIMITATION AND FUTURE DIRECTION

Although the use of LLM agents in financial trading has achieved many successes, limitations still exist in current research.

From an architectural perspective, most agents rely on closed-source models (e.g., GPT-3.5/GPT-4), which raises concerns about data privacy and restricts the ability to customize model development. Additionally, our review reveals that most studies applies LLMs through in-context learning without any fine-tuning, with only [19] tunes the LLM during training. The effectiveness of fine-tuning LLMs for trading agents remains an open question. Another significant issue is the inference latency, which can be a bottleneck, making these models impractical for high-frequency trading. Moreover, integration with existing trading systems is rarely discussed in the literature we surveyed.

From a data perspective, while agents typically use textual data such as news and fundamental data, few utilize social media data, which can significantly influence financial market (i.e. the Game Stop Short Squeeze [33]).

From an evaluation perspective, backtesting is predominantly confined to the US and Chinese stock markets, with notable absences in other financial markets such as derivatives, bonds, or commodities. Additionally, backtesting periods are generally short, and few studies consider trading costs. Expanding evaluations to include these other markets and accounting for trading costs could unveil new opportunities, particularly given the potential sensitivity to news data in these markets.

Lastly, agents with different trading style or personality tends to perform differently in making trading decisions. However, few studies have conducted ablation studies to explore the underlying reasoning processes of LLMs in their trading decisions. Utilizing simulated environments could be a promising approach to gain deeper insights into the LLMs' decision-making processes and patterns.

6 CONCLUSION

In this survey, we systematically reviewed all relevant works that leverage LLMs as trading agents, focusing on their architectural design, data inputs, and evaluation methods. Although this is an emerging field with relatively few studies to date, we found that LLM-powered trading agents demonstrate significant potential in extracting signals from massive amount of textual information and making informed decisions. However, there are still challenges including the reliance on closed-source models and the integration issues with existing trading systems and human traders, which can be important directions for future research.

REFERENCES

[1] Siyu An, Qin Li, Junru Lu, Di Yin, and Xing Sun. 2024. FinVerse: An Autonomous Agent System for Versatile Financial Analysis. arXiv:2406.06379 [cs.CE] https://arxiv.org/abs/2406.06379

[2] Clifford S Asness, Tobias J Moskowitz, and Lasse Heje Pedersen. 2013. Value and Momentum Everywhere. *Journal of Finance* 68, 3 (2013), 929–985. https://doi.org/10.2139/ssrn.2174501

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen Technical Report. arXiv:2309.16609 [cs.CL] https://arxiv.org/abs/2309.16609

[4] Vitalik Buterin and Ethereum Foundation. 2024. Ethereum: A Decentralized Platform. https://ethereum.org/en/ Accessed: 2024-07-10.

[5] China Securities Index Company. 2024. CSI 300 Index. http://www.csindex.com.cn/en/indices/index-detail/000300 Accessed: 2024-07-10.

[6] Marvin M Chun and Marcia K Johnson. 2011. Memory: enduring traces of perceptual and reflective attention. *Neuron* 72, 4 (Nov 2011), 520–535. https://doi.org/10.1016/j.neuron.2011.10.026

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL] https://arxiv.org/abs/1810.04805

[8] Yujie Ding, Shuai Jia, Tianyi Ma, Bingcheng Mao, Xiuze Zhou, Liuliu Li, and Dongming Han. 2023. Integrating Stock Features and Global Information via Large Language Models for Enhanced Stock Return Prediction. arXiv:2310.05627 [cs.CL] https://arxiv.org/abs/2310.05627

[9] Encyclopaedia Britannica, Inc. 2023. MACD (Moving Average Convergence Divergence). https://www.britannica.com/money/macd-moving-average-convergence-divergence Accessed: 2024-07-20.

[10] Georgios Fatouros, Konstantinos Metaxas, John Soldatos, and Dimosthenis Kyriazis. 2024. Can Large Language Models Beat Wall Street? Unveiling the Potential of AI in Stock Selection. arXiv:2401.03737 [q-fin.CP] https://arxiv.org/abs/2401.03737

[11] Fidelity. [n. d.]. Relative Strength Index (RSI). https://www.fidelity.com/learning-center/trading-investing/technical-analysis/technical-indicator-guide/RSI

[12] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring Mathematical Problem Solving With the MATH Dataset. arXiv:2103.03874 [cs.LG] https://arxiv.org/abs/2103.03874

[13] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (nov 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[14] Dennis Huynh, Garrett Audet, Nikolay Alabi, and Yuan Tian. 2021. Stock Price Prediction Leveraging Reddit: The Role of Trust Filter and Sliding Window. In *2021 IEEE International Conference on Big Data (Big Data).* 1054–1060. https://doi.org/10.1109/BigData52589.2021.9671412