香港中文大學
The Chinese University of Hong Kong

G factor portfolios by LLMs in Chinese stock market

FTEC5910/5920: Fintech Idustrial Project

April 4th, 2025

Wenbo Zan 1155217739

Contents

1 Major Concept Background

1.1 Algorithm Trading and Frequency Improvement

Algorithmic trading is an important innovation in the field of finance, which take in advantage of computational models to make trading decisions. By the definition given, algorithmic trading specifically refers to the use of computer programs to automate one or more stages of the trading process in electronic financial markets [4].

To further understand what is algorithm trading, we can divide the perspectives of it and we can pay attention to sub listed aspects: pre-trade analysis (data analysis), trading signal generation (buy and sell recommendations), and trade execution. Each stage of this trading process can be conducted by humans, by humans and algorithms, or fully by algorithms. Or to simply generate as: Making decisions by objectively or subjectively. By the help of algorithm trading, every movement done can have a clear reference, avoid the chance to make irrational behavior, thus a strong protection at the value at risk.

More than than that, in today's complex and volatile financial markets, algorithmic trading enables trading operations to be carried out efficiently, accurately and in an emotionally independent manner. The development of nowadays' calculation ability obviously utilized data science skills to manage tons of data and make quicker decisions. Larger amounts of market data, more rapidly can computers identify potential trading opportunities and execute trading orders based on predefined strategies, comparing to human judgment.

With more and more options relate to algorithm trading, making this method grows to a big family of measurements, portfolio managers are expecting to keep improving their analyzing skills, which ensures to deal with more specific data, indeed making wise decision, since they can track as more information as they can. In this background situation, trading and analyze frequency is developing, hence we tend to introduce high-frequency trading.

The definition for high-frequency trading is various, and in general, we can refer this idea by: A specific branch of algorithmic trading, which pays more attention on speed and precision. The most agreed-upon characteristic of HFT is its use of high-speed, sophisticated computer programs (SEC, CESR, MiFID II) or strategies (NAFM, ASIC) to generate, route, and execute orders[12]. Normally in a very short period of time, these portfolio shall deal with high-level data, for example, tick by tick, and make quick response. High-frequency quantitative trading can even execute trades at millisecond or microsecond speeds at most, aiming to profit from small price movements in the market.

In this article, we tend to set the benchmark for high-frequency as a perspective to have comparison, thus analyze our experiments both in daily and minute-level data, trying to observe excess earnings in data difference and sentiment efficiency, aiming to find the best way to optimize our portfolio in a time-related factor.

1.2 ESG-Information Influence

Today, ESG is critical to our investment decisions. However, emerging trend suggests that a more detailed perspective analysis in specific markets plays a more fundamental role. For example, within the context of Chinese listed companies, corporate governance (G perspective in ESG analysis) has gained a unique influence. In the past, some Chinese listed companies have been exposed to problems such as loose management structures, lack of internal monitoring mechanisms, and unscientific decision-making processes, which have led to many operational disruptions, not only affecting their own stable development, but also ked investors to many uncertainty risks. Based on this background, corporate governance has begun to be highly valued since market improvement. A good corporate governance system promotes orderly internal operations, good interaction with external stakeholders and government regulation.

Since Governance factor became a key in affecting enterprises' achievement about sustainable development and win the favor of investors in the capital market, from the perspective of safeguarding the healthy and stable development of enterprises, continuously strengthening the level of corporate governance (G) is an inevitable choice for those listed ones. Therefore, obtaining this information for investment strategy analysis is a fundamental measure in order to gain the market's trust, and take advantage of it in the company's development, which is positively proportional in terms of return analysis. More than that, some points about Governance such as compliance with regulations and make optimizations can obviously reduce market risk by reducing unnecessary losses and gaining market traction while complying with the regulatory climate. Previous jobs has proved using the existing data, which we can have a comparison on. It emphasizes that: Regression analysis can prove that there was a significant relationship between stock returns and corporate governance practices in NSE (Nairobi securities exchange) market, companies practicing good corporate governance practices are likely to enjoy higher stock returns [6]. Similar as Chinese A share stocks, the base idea from NSE has conclude the importance of making use of Governance factor, for our further studies to reflect on in Chinese A shares data to prove its efficiency.

1.3 Natural Language Processing and Large Language Model

Natural Language Processing(NLP), is a way to deal with natural characters for machines to learn from and get used to. Normally, when we deal with NLP problems, we always refer to the functions done by neural network mechanism, as it has been proved more efficient in dealing with multiple words and sentence recognition. Taking a path of the development of this method, it is initially giving out by Bengio in his article *Neural net language models* [1] , the specific definition is stated as: A neural network language model is a language model based on Neural Networks , exploiting their ability to learn distributed representations to reduce the impact of the curse of dimensionality.

With the development of neural network methods, more skills to improve accuracy of NLP need was introduced. In 2013, the word representation method took the representative role, which enhanced the sparse vectorization idea to conclude words difference. In this area, Word2vec [3] was introduced as a library for English word reference. Later in 2014, Recurrent Neural Network (RNN) methods such as Long Short Term Memory (LSTM), Gated Recurrent Unit (GRU) have been taken in the improvement of NLP recognition. Further more, the attention mechanism was developed to focus on the connection between words, forming the idea of sentences and paragraphs, which highly advanced the accuracy of the NLP projects.

In 2017, the base methodology for token recognition and response model, the Transformer Network, was announced in NIPS [8]. It is so important that after its issue date till now, most of the response and recognition models was developed base its structure. Even after OpenAI issued the CHATGPT series [5], it still use this structure to develop its products such as ChatGPT 3.5/4 and DALL-E platform. It is fortunate that the NLP have a lot in common with Large Language Model construction, since they all based on the tokenization process as the deep base and move to different perspective. In words, NLP projects tend to words recognition, while LLMs tend to resource allocation to generate response. For example, the Llama model [7] gives us basic LLM structure and allowed us to train the model using our own data to take advantage of the fine-tuning algorithm.

2 Project Motivation and Literature Review

All three aspects in the background statement can play effective roles in financial decision analysis, so how to combine these three is of great significance in our design, which can help us build a new type of underlying screening mechanism. Firstly, high-frequency trading provides efficient trade execution, ESG information provides a more comprehensive evaluation perspective, and large language models provide powerful analyze support. The combination of them can theoretically provide a richer and more powerful tool for decision-making. Secondly, we can observe the advantages in terms of analysis speed, since the characteristics of high-frequency trading, neural networks and the efficient computational power of large language models synergies with each other, which can greatly improve the speed of financial analysis and enable investors to respond more quickly to market changes. Finally, in terms of analysis material, ESG information enriches the content of analysis, with NLP's ability to process various types of data, providing a more comprehensive source of information for sentiment catching. In summary, the combination of the three can undoubtedly bring greater advantages to financial decision-making, indeed gives solid motivation to conduct a study on the idea.

There are works done to combine such ideas within two aspects of those three, which we can have some reflections on. In the work done by Xinli Yu et al. [11], the existing LLMs such as ChatGPT-4 can make a well-thought decision by reasoning over information from both textual news and price time series and extracting insights, leveraging cross-sequence information, and utilizing the inherent knowledge embedded within. Also it stated that available basic LLM such as Open-LLaMA, after fine-tuning, can comprehend the instruction to generate explainable forecasts and achieve reasonable performance, albeit relatively inferior. This work is done by NASDAQ-100 stocks, which proved the ability of combining financial date within existing response models. More than that, in Financial Sentiment Analysis (FSA) area, existing idea have been proven successful by adopting Minsky's theory of mind and emotions into consideration rather than fine-tuning [9]. More than these, it is a trend in today's financial markets worldwide to take AI development as their own progress, and indeed has created lots types of LLMs related to financial area. This is helpful for us to make a reflection on which they have done great jobs reaching rarely good precision, still we want in deeper perspective to combine three factors together to leverage their uniqueness and gives a new road.

3 Problems Description

With existing work have done, the problem is pretty clear, requiring me to extract useful Governance factor information from news sentences, to generate ratings by LLMs chosen and build up an sentiment-combination quantitative strategy. There are three key perspectives that needed to be considered, which as following:

3.1 LLM Efficiency

Due to the popularity of LLMs that many different types of them have emerged on the market, each of which uses different training data and architectures, which poses a challenge in summarizing the conclusions in our experiments, and thus we need to conduct multiple sets of controlled experiments to test the validity of the extant models in G factor analysis. This will not only help us to select the most suitable model for our factor analysis, but also provide directions for further improvement and optimization of the model choices and even more a benchmark for further fine-tuning process. At the same time, obtaining different sentiment factor scores from a wider variety of language models can also tell us what preferences each type of language model has in processing

Governance information, which can also help us to lay the foundation for further research requirements.

3.2 Time Effectiveness of Information

Over time, information may lose its value as new situations and data emerge. To better understand the timeliness of information, we need to compare different frequencies of transactions. We need high frequency trading to be compared with traditional low frequency trading strategies to consider their realization of information validity on the same disclosure, thus arguing for the strength of high frequency trading. We plan to observe the extent to which information is profitable on different time scales by comparing trades of different frequencies. In this way, we can better understand the value and impact of information on different time scales and thus make more informed decisions.

3.3 Outcome Expected

In the result expectation, we aim to develop a Generative Portfolio Analysis Factors Chart based on a multi-model framework to assess the effectiveness of generated scores using quantitative methods. The analysis will include a profit and loss line chart, visually demonstrating the model's performance and highlighting any outper-formance trends. Beyond visualization, this analysis will serve as a direct evaluation tool, measuring how well different models generate meaningful scores from the given data. Key performance indicators such as Sharpe ratio, maximum drawdown, and volatility will be incorporated to ensure a comprehensive assessment. This structured approach will not only validate the generated scores but also help refine and optimize model performance, ensuring their reliability in investment advice. Apart from the basis model performance, we will also done our investigations on the news filtering and score similaity analysis to fully aware of the data we have in hand.

4 Methodology Design

The pipeline for existing work shall work as the below Figure 1. The total procedure will be divided into two parts, part 1 mainly focus on the part in red line circle, which is the data processing work, and the rest being the experiment procedure, which consist of the backtest and analysis.

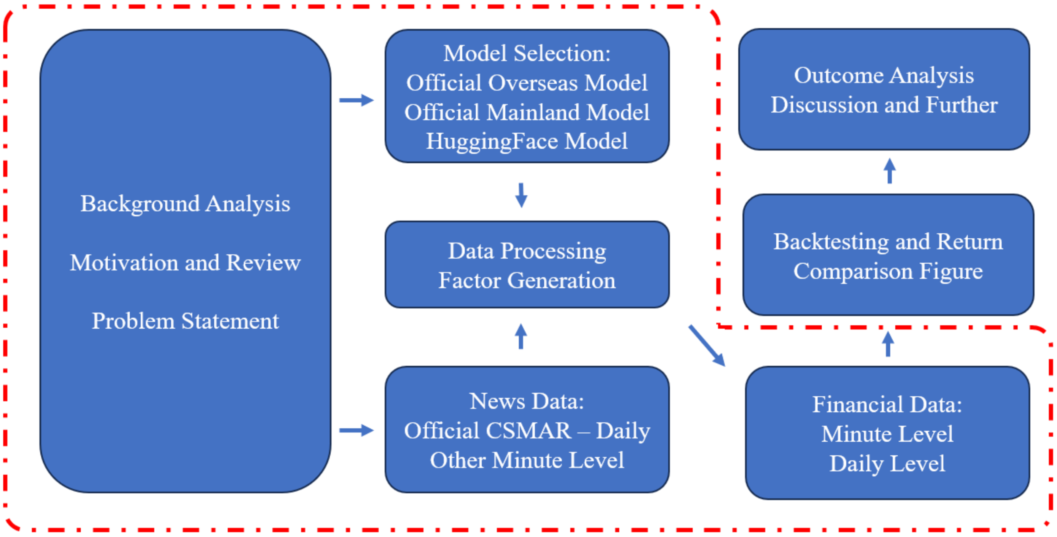


Figure 1: Methodology Pipeline Figure

5 Data Detail and Tool Preparation

According to the figure above, the information we need to prepare mainly comes from three perspectives: Financial Data, News data, Models. Here we will introduce the data source origins and how we will process each data to accomplish our goals.

5.1 Data Collection and Comparison

5.1.1 Direct Method to obtain ESG news

I need to obtain the real-time news as raw samples for further processing to put in model analysis. Although there are lots of news sources in the market, they need to be paid or the number of news able to be achieved is greatly limited. Here we can obverse from **AKshare** open-source Platform, and also **Ricequant** forum with existing ESG rating data as:

- **1: Sina Finance - ESG Rating Center - ESG Rating detail - Rating Data**
- **2: Sina Finance - ESG Rating Center - ESG Rating detail - MSCI**

- **3: Sina Finance - ESG Rating Center - ESG Rating detail - Refinitiv**
- **4: Sina Finance - ESG Rating Center - ESG Rating detail - Yiding**
- **5: Sina Finance - ESG Rating Center - ESG Rating detail - chindices**
- **6: Ricequant - ESG Rating**

These open-source can serve as a reflection after the rating of the LLMs given to check the alliance of the data with the official rating table. Those well-known quantitative trading platform and provides professional ideas about the esg news but without a judging logic as well as the basic idea of rating and the timestamp which we can not take in our experiments.

5.1.2 Direct News sentiment Data

There are multiple Database online for our reference, but few will suits our need, which have the news content, the timestamp and related companies ID. Here we can set categorizations in those Databases and I will give an explanation of my work:

Proper Data Type with no Access:

Database listed shall be those without chance to catch but good in shape

- **CNopendata**: CNopendata is a comprehensive data platform. It has rich data resources, including patent data, industrial and commercial registered enterprise data, listed company data and other ten data series. Its data volume is large and diverse, with numerical and textual data co-existing. However it is unable to obtain long-term data, which need to contact for cooperation.
- **Tushare**: Tushare is a powerful financial data interface platform. It covers a wide range of financial data such as stocks, funds, bonds, etc., providing users with a wide range of data options. Data acquisition is very convenient, users only need to call the interface through the code, it is easy to get the required data, which greatly improves the work efficiency. However the download speed is too slow (400/m) and require payment, which we shall ignore.
- **Ricequant**: Ricequant is a quantitative investment platform which has rich financial data resources covering a wide range of financial products such as stocks and

futures. The data is updated in a timely manner, which can provide strong support for quantitative investment strategy formulation. Ricequant's data is complete and organized, but it lacks news body and news summary sections. The advantage is the existing data that has been cleaned, but the disadvantage is that the time span of the cleaned data is relatively small, making it difficult to organize.

- **Datayes**: Datayes is a professional data service platform. It gathers massive and diverse financial market data, including data on various assets such as stocks, bonds, futures, etc., providing ample materials for quantitative investment, investment research analysis, and more. It has strong data processing capabilities, which can accurately process and deeply integrate data. However, the storage method requires server storage, too expensive for me to afford.

Wrong Data Type with Access:
Database listed shall be those can be acceptable but hard to process:

- **Datago**: Datago is a platform dedicated to data services. It has rich and high-quality data resources covering multiple fields such as finance and economics, with detailed and comprehensive data collection and organization. However, the data permission only applies to news data related to newspapers and magazines, with a relatively small quantity, and the news ID cannot directly correspond to the news data. No textual data was found here.

- **Opendatalab**: Opendatalab is a platform dedicated to open data. It gathers massive and diverse data resources, covering numerous fields, and can meet the data exploration needs of users in different industries. The news information with the highest frequency of information is provided here, but it is difficult to clean without corresponding information from the company. The disadvantage is that there is no classification of the correlation between listed stocks and no data cleaning. The news time is not precise to the second level either.

In the final, I choose the Eastmoney platform, which is the biggest financial data provider in China that has most adequate real-time news that I could utilize for analysis. In this method I utilized AKshare to grab minute level news data from Eastmoney website, to introduced as the data need for minute-level strategies.

From a daily perspective, we will utilize the full dataset rather than limiting ourselves to minute-level data we choose, which only covers specific transaction periods. This broader approach provides clearer insights and allows for a more comprehensive comparison of data variations. By leveraging daily-level data, our analysis will be more accurate and reliable, ensuring a professional and well-rounded experimental outcome.

5.2 Model Selection

For the LLM choices, we shall have the listed as our targets, which we finally choose to have done the experiments based on the current limitations(the capability of the computational power). We have selected those as our target to test:

- **Ernie-Lite-8k**: The *Ernie-Lite-8k* is a lightweight powerful language model developed by Baidu, optimized for efficiency and scalability within an 8K context window. The lite meabs it is designed for applications requiring rapid inference and lower computational overhead, but retains robust performance in tasks. Its streamlined architecture makes it ideal for edge devices or scenarios where resource constraints preclude larger models.

- **Deepseek-R1-7B**: The *DeepSeek-R1-7B* is a distilled version of the larger DeepSeek-R1 series, designed for efficient local deployment on devices with limited computational resources, which we here use Ollama platform to deploy. The distilled model is enabling offline use in scenarios such as document processing and content generation. As a root production compare to the "full-blooded" 671B version, it is notably in some research claims, that specialized 7B models can rival the 671B version in niche tasks. Here we are interested in its special structure of "Deep Thinking", and want to have the experiments on its performance.

- **Llama-3.2-3B**: The *Llama-3.2-3B* is Meta's efficient 3-billion-parameter model from the Llama 3.2 series, optimized for on-device AI with a 128K context window. The reason we choose it shall be it can outperform similar-sized models like Gemma-2B and Phi-3-mini in tasks like multilingual dialogue and summarization while supporting quantization (QLoRA/SpinQuant) for faster, lower-memory mobile deployment. Ideal for edge AI, it balances performance and efficiency in local inference scenarios. More than that, Llama will serve as a fine-tune base model for us to judge the information abstraction in the base point.

- **Spark-Lite**: The *Spark-Lite* is a compact version of Spark series by Spark Platform, optimized for efficient deployment on edge devices and mobile platforms. It delivers fast inference speeds while maintaining strong performance in conversational AI and text processing tasks, making it ideal for lightweight applications.

- **FinBERT-2019**: The *FinBERT-2019* is a pioneering financial domain-specific language model based on BERT, introduced in 2019 as one of the earliest adaptations of BERT for finance. It was trained on a large corpus of financial texts, including news, reports, and company filings, to enhance performance in financial sentiment analysis, entity recognition, and other NLP tasks in the financial sector. The finbert is a base model for sentiment analysis which it shall be use the input paragraphs to judge the sentiment in 3 options.

- **Finbert-HKUST**: The *FinBERT-HKUST* is a specialized financial language model developed by HKUST, tailored for financial text analysis, including sentiment analysis, entity recognition, and market prediction tasks. It builds on the BERT architecture but is fine-tuned on financial corpora like earnings reports, SEC filings, and financial news to capture domain-specific nuances. Similarly as FinBERT-2019, it also have 3 options to choose in sentiment analysis. However, it is rather different since it is a masked LLm for users to check on the answer based on the given prompt and specific mask, which make it useful in generation.

- **DISC-FinLLM**: The *DISC-FinLLM* is a finance-specialized LLM developed by Fudan University, fine-tuned from Baichuan-13B-Chat with LoRA adapters for financial consulting, document analysis, accounting, and news interpretation. It features a modular design with four task-specific adapters that can be switched without reloading, supports RAG for real-time market data integration, and was trained on 250K high-quality financial instructions (DISC-Fin-SFT). The model and dataset are open-sourced, with a live demo available for testing.

- **Tongyi-Finance-14B**: The *Tongyi-Finance-14B* is Alibaba's financial domain-specialized large language model, offering 14 billion parameters optimized. Built on

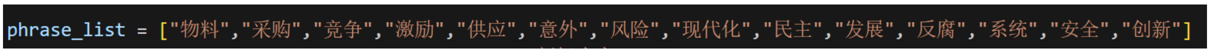
Qianwen architecture, it demonstrates superior performance compared to general-purpose models. The model incorporates real-time financial data processing capabilities while maintaining robust Chinese-language support, making it particularly effective for China's financial markets.

These models ranges from official models released by large companies from overseas and mainland, the basic model without fine-tuning, and fine-tuned model using specific financial data. In this perspective, we can observe the difference in models about factor generation. Still, after the score generation test about the news, under the simialr prompt to make fewer difference, 4 listed become the final choice of this test, with their name: *Ernie-Lite-8k*, *Deepseek-R1-7B*, *Llama-3.2-3B*, and *Spark-Lite*. The specialized LLMs are not chosen mostly because the response structure were too bad to process, or lack of most of the answer, which only few will be considered useful, thus we could say the casual model may be more accurate in this project.

5.3 Data Processing

5.3.1 News Categorization

At this stage, there should be a designed code that can determine whether a paragraph belongs to the required content based on the keywords found in the search, and then return it to the stored list information, here by notice, we checked amount of articles to generate the words table as the following picture. This operation can significantly reduce the amount of data that needs to be processed while filtering effective data to prevent data noise from affecting the experimental results. Hence the choice of words may not be that much and the word choice should be short enough in order to select news precisely as we can, with 1/10 of the raw data will be chosen as the final check for models to judge twice.



```
phrase_list = ["物料","采购","竞争","激励","供应","意外","风险","现代化","民主","发展","反腐","系统","安全","创新"]
```

Figure 2: Word list

5.3.2 Financial Data Types

The financial data is easy to be obtained, however, in order to stay the same in data without difference caused by companies behavior, we tend to use minuet level data as both of the minute-level experiment and daily-level experiments, and take in use of the final minute of a day to represent the daily data. The data shape will be in shape as Figure 5.

5.4 Model Response Design

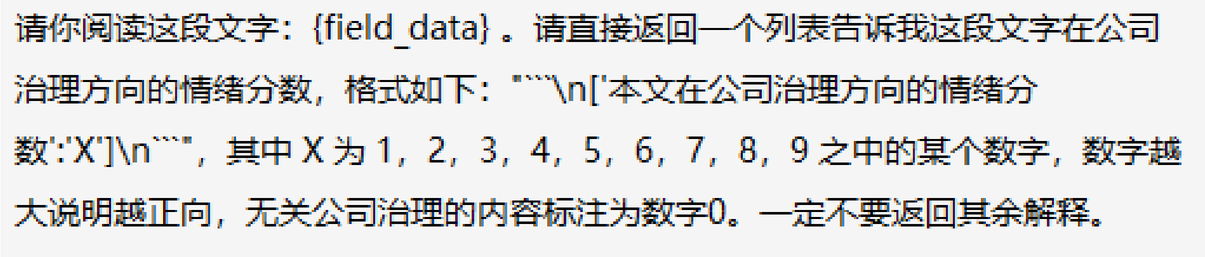
5.4.1 Factor Rating Methodology

In the model design after we have selected, and after the shuffling of the news data, we tend to introduce the sentences we obtained into the large language models to gives the rating as set in the Figure 6. In order to receive this data type we will introduce the prompt setting in the next section, which will be shown in yellow. Also, since the model down by different dataset will have different language setting, here we created the prompt in both language, English and Chinese which they share the same meaning and sturcutre construction, also shown in the following figure.

5.4.2 Prompt Base and Generation result

Listed shall be the prompt setting that will be put for LLMs to answer.

> Please read this article:***field-data***. Please directly return a list to tell me the sentiment score in Governance factor of the given article, with a solid format as: ”['The sentiment score in governance factor':'X']”, here X is a number in 1,2,3,4,5,6,7,8,9, larger the number, more postivive the sentiment, and if the article is no related to the governance factor, the X is 0. Do not return any other word explanations



请你阅读这段文字：{field_data}。请直接返回一个列表告诉我这段文字在公司治理方向的情绪分数，格式如下："```\n['本文在公司治理方向的情绪分数':'X']\n```"，其中 X 为 1，2，3，4，5，6，7，8，9 之中的某个数字，数字越大说明越正向，无关公司治理的内容标注为数字0。一定不要返回其余解释。

6 Backtest Methodology and evaluation matrix

After the Basic introduction on the model and data choice, we shall turn to how to generate the data that we need. Here by notice, since we illustrate our news data by order into the models, we will store the data into a list by order into a json file in the consideration of easy reading, and make further adjustments. Some pre notice should be given that, for all the models we finally choose, the backtesting procedure construction should be the same as our goal is to check the ability of each model and thus we have set the same benchmark choice in the certain period and exact the same grading policy. The final outcome is to compare the portfolio performance with a one-day turnover and one-minute turnover, which by one of my goal - to check the time effectiveness of paragraphs in the same period but different execution time.

The Chinese stock market has a lot specific rules, especially some may cause impact on our model. For example, the transaction cost will make huge impact on the high-frequency analysis since the large amount of the orders may cause multiple cost which will affect the outcome tremendously. Other than that, the market regulations also make a hint that only T+1 transactions will be offered in the Chinese market, which indicate that our trading methodology in minute level share no real-market value. However, based on the investigation need for information gap listed in the question aspect above, the T+0 transaction of minute level stock portfolio should also be researched.

6.1 Final score data shape

In this section, we introduce the final structure of the score data, building upon our prior understanding of the score generation process. As established, the procedure produces scores within a numerical range from 0 to 9, which such systematic approach guarantees every data entry being associated with a clearly defined score. The final DataFrame will largely retain the format of the initial one with news paragraphs, but as the key distinction being the addition of several new columns. Each of these columns will be named according to the respective model responsible for generating the scores. This enriched structure allows for an intuitive and efficient comparison of different models' outputs within the same framework, making it easier to draw insights and evaluate performance.

Additionally, we have rigorously checked the stock index structure, confirming its perfect alignment with the index names in the price data, and adding a column to store the checking outcome to ensure the data availability. This alignment plays a crucial role in maintaining data consistency, preventing potential mismatches, and ensuring smooth integration in subsequent analytical steps. By establishing a well-structured and coherent dataset, we set a solid foundation for the methodological framework that follows, enabling a more refined, accurate, and insightful analysis of the data.

6.2 Backtest Methodology

The minute and daily return shall be share a similar methodology which will be as: both of strategies are choosing a specified range of the stocks, which is CSI500 stocks to work as the base, which also we will use the CSI500 index as our portfolio benchmark to calculate the wining rate for our strategy. The calculation methodology shall be as the next period close price for a certain stock divided by the current period close price for this stock and subtract 1 to get the return, which the hidden logic shall be that we will buy this stock this period of time base on the level of the data and then sell it on the next period. The calculation of this logic is as follow:

$$\text{Return} = \frac{\text{Close Price}_{\text{next}}}{\text{Close Price}_{\text{current}}} - 1 \tag{1}$$

In the formula represents the stock close price for the next period and the current period. This is the base for each item in the matrix of generation of the pivot return matrix by the time and the code index. It is pretty clear to understand, and easy for later judgment. Besides that, we need a weight matrix also to conduct the final return of the period portfolio, which is from the key of this project, the grading result. Since we have introduced the grading of the scores above, we can then generate the corresponding weight matrix similar as the return table to take advantage of and then by dot plot to have the return in each period of time and code. The generation of the weight is also simple, as we can use the adding up of the total score to represent the outcome weight and the divide by the total score of that period to have the final weight, which in the expression as:

$$W_{\text{specific}} = \frac{S_j}{\sum_{i=1}^{n} S_i} \tag{2}$$

The generation of the total score for single stock shall have a limit that only score than 5 will be considered positive and the being selected. All by that, we can simply adding up the total return to have the final one, which we can have the mathematical equation for understanding as:

$$R_{\text{total-specific-time}} = \sum_{i=1}^{n} w_i \cdot R_i \tag{3}$$

Now, I have the matrix of picked stocks and with their portfolio return for every trading period, which we have taking in use of the original score matrix. The profit and loss will base on the period outcome to use the cumulative return method and will generate the profit and loss line base on the equation result stored in a list. Then we can judge their corresponding result in the following evaluation process.

6.3 Evaluation Strategy

6.3.1 Scores data evaluation

We have verified the score generation process, with results shown in the two accompanying figures. Both charts display the distribution of scores across all evaluated models, revealing strikingly similar patterns in their shapes. This consistency indicates that the grading produces comparable outcomes for different models, suggesting our evaluation framework is both fair and reliable.

The first figure (Figure 3) presents the total scores, while the second (Figure 4) breaks them down into finer detail. Despite the difference in scale, their nearly identical trends confirm that the scoring system captures meaningful performance differences rather than random variations.

This observation is further supported by the numerical analysis in Figure 7, where we examine the exact score distributions. The number shall share similar multiple amount, which indicate similar shape.

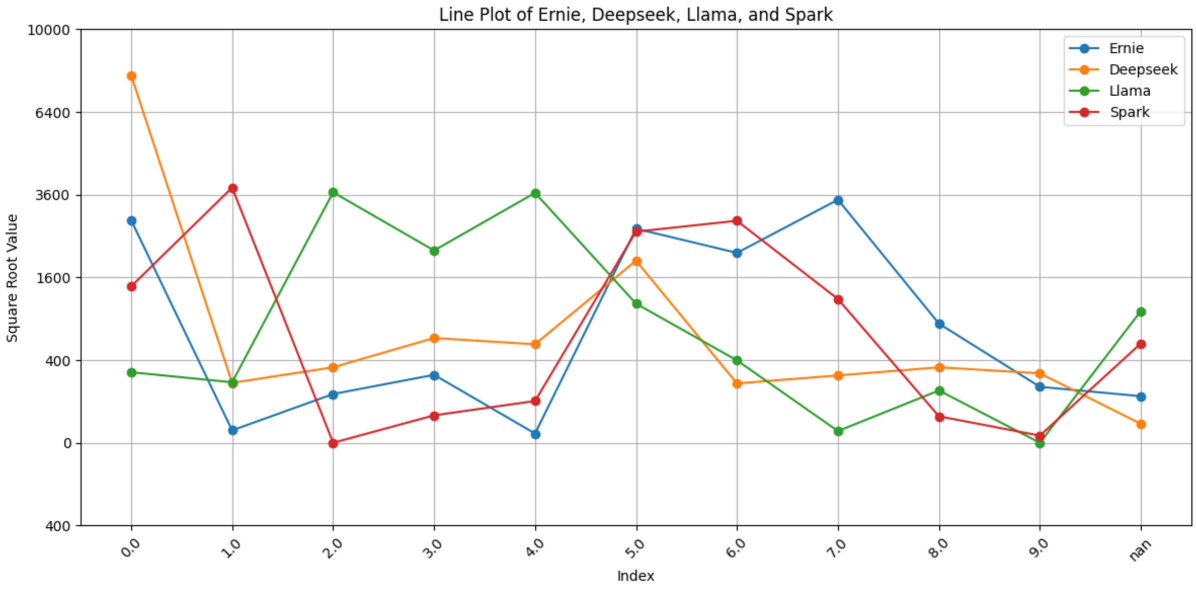


Figure 3: Total socre

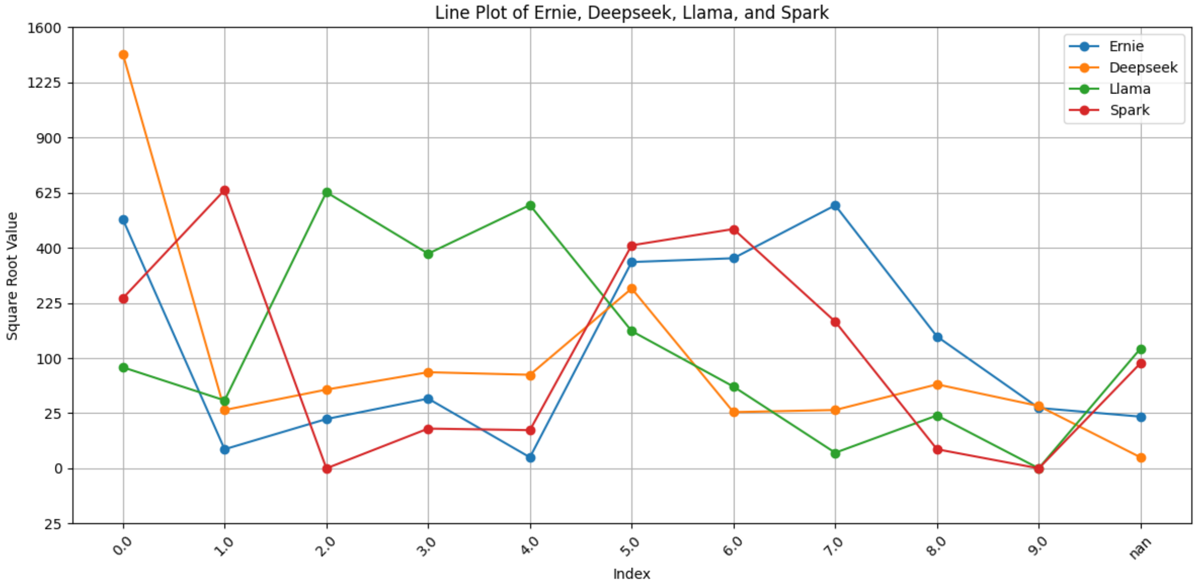


Figure 4: Minute score

6.3.2 Backtest result evaluation

The strategy's effectiveness is evaluated through five key indicators:

- Maximum Drawdown – Measures worst-case capital loss.
- Sharpe Ratio – Quantifies risk-adjusted returns.
- Win Rate – Trades the frequency of profitable trades.
- Annualized Return – Annualized profitability.

- Volatility – Reflects return variability and risk exposure.

Together, these metrics provide a rigorous assessment of both risk and return dynamics. These carefully selected indicators provide a multidimensional assessment of risk and return characteristics, offering a holistic view of the investment strategy's effectiveness.

Maximum drawdown (MDD) identifies the largest peak-to-trough decline in profit return table over a throughout period, offering a clear view of potential capital erosion. MDD directly reflects the strategy's worst-case loss scenario. The mathematical equation shall be as following:

$$MDD = \max_{1 \leq i \leq j \leq T} \left(\frac{V_i - V_j}{V_i} \right) \tag{4}$$

Where in this formula, the Vi represent the value peak in time i, with the corresponding trough in time j. The time T will be the time length for our experiment, which hear is 28 transaction day from 2024/11/18 to 2024/12/25.

The Sharpe ratio evaluates excess return per unit of risk, with higher values indicating superior risk-adjusted performance, the compression for both profit and risk made it the key consideration which is widely used in most investment analysis. In the formula base, it shall be written as:

$$SR = \frac{E[R_p - R_f]}{\sigma_p} \tag{5}$$

The formula includes the portfolio return (Rp) and the risk-free rate (Rf), which Rf is non applicable here in our experiment since the portfolio analysis based on the true return itself. The sigma indicate the standard deviation of the portfolio here, as well as the risk. The higher the Sharpe value, the more you shall earn in this portfolio rather in same risk measurement.

The win rate measures the proportion of periods in which a strategy's returns outperform the benchmark's returns, excluding cases where the strategy's return is zero (indicating no trading activity). First, we filter out all periods where the strategy's return is zero to focus only on meaningful comparisons. Then, for each remaining period, we check whether the strategy's return exceeds the benchmark's return. The win rate is computed as the average of these binary outcomes (1 for outperformance, 0 otherwise), yielding a value between 0 and 1 that represents the fraction of times the strategy beats the benchmark.

$$WR = \frac{1}{N} \sum_{i=1}^{N} I(S_i > B_i) \tag{6}$$

In this formula, we know that where Si and Bi are the strategy and benchmark returns in period is the number of valid (non-zero strategy return) periods, and I() is the indicator function (1 if true, 0 otherwise).

The annualized return is computed by first determining the total cumulative return of the strategy over the entire period, calculated as the product of (1+returns) minus one, which accounts for compounding effects. This cumulative return is then annualized by raising it to the power of the ratio of the total number of trading periods in a year (e.g., 252 days for daily data or 252 × 240 minutes for intraday data) to the actual number of periods in the returns data. This scaling converts the cumulative return into an equivalent annual rate, allowing for consistent comparison across different time horizons. The formula shall be as following:

$$R_{\text{annualized}} = \left(1 + \prod_{t=1}^{T}(1 + r_t) - 1\right)^{\frac{N}{T}} - 1 \tag{7}$$

where in this equation rt is the return at time t, T is the total number of periods in the returns data, and N is the number of periods in a year.

The volatility calculation measures the dispersion of returns, representing the degree of variation in a trading strategy's performance over time. This function computes volatility as the standard deviation of returns, which quantifies how much the returns typically deviate from their mean value. A higher standard deviation indicates greater risk and price fluctuation, and vise versa. The calculation uses the sample standard deviation by default, which provides an unbiased estimate of the underlying population volatility when applied to historical return data. The formula shall be as listed:

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (r_i - \bar{r})^2} \tag{8}$$

where $sigma$ is the volatility, N is the number of observations, ri represents individual returns, and *rbar* is the mean return.

By systematically analyzing these key metrics derived from market data, we gain a comprehensive framework for evaluating trading strategy performance. Each indicator provides unique insights into different aspects of portfolio behavior - from risk exposure to return patterns - creating a multidimensional perspective essential for robust performance assessment. Together, these quantitative measures form the foundation for our detailed analytical examination, enabling data-driven decision-making and strategy optimization.

7 Result analysis and discussion

After the introduction which about the technics, here we will analyze and discuss the comparative results between two distinct time periods, highlighting the differences and significance of daily and minute-level investment decision-making, as well as checking of the various models ability. Here should by hint that beyond theoretical scenarios, we need incorporate realistic transaction costs to validate our strategy and assess its practical performance. The total transaction costs in our analysis consist of three key components: Stamp duty, Brokerage commission fees, Securities transfer fees. Which can be in total as 0.00641 percent in buy, and 0.05641 percent in sell. These cost factors are systematically incorporated into our evaluation framework through the following equation formulation:

$$R - costed = R * 0.99937184 \tag{9}$$

Where the subtraction number shall be calculated as (1-0.0000641) / (1+0.005641) . Since the calculation procedure shall be clear, we can simply done the cost factor analysis by the math division.

7.1 Minute-level analysis

Firstly, we shall see from the given total pnl line in the Figure 8. The provided PNL curve illustrates the performance of four different models—Ernie, Deepseek, Lama, and Spark Benchmark—without accounting for transaction costs. The data reveals significant variations in performance among the models, with Ernie and Deepseek showing negative returns, while Lama and Spark exhibit positive returns. However, four models all outperform Benchmark. Spark stands out as the top performer with a PNL of approximately 0.04, indicating a relatively strong compounding effect over the observed period. Llama follows with a modest but still positive PNL of around 0.02, suggesting it is profitable but less so compared to Ernie.

On the other hand, Ernie and Deepseek shall be less in profit or to say even in loss. These negative values imply that these models would incur losses if deployed even in a without-cost trading scenario. The stark contrast between the top and bottom performers highlights the importance of model selection, as the choice between Ernie and Spark, for example, could lead to a 5 percent swing in performance.

Overall, the data underscores the critical role of model efficacy in generating positive returns. While Lama and Spark demonstrate viability, Ernie and Deepseek would require substantial improvements or alternative strategies to avoid losses. The exclusion of transaction costs in this analysis means the actual performance gap could be even wider, as costs would further erode the margins of the already underperforming models.

After a throughout visual of the comparison, we shall divide into the model single chart. Firstly we shall take a look at the circumstances with the cost. The Figure 9, Figure 10, Figure 11, Figure 12 are the pnl line separately for four models with the transaction cost adding into consideration.

We can observe from the data that all the pnl lines are showing a trend of down-sloping, since the adding of the transaction in minute level will cost too much in the transaction, and thus making a huge drop in the profit, which cost down sloping type. But rather than just down sloping, we can observe that in these charts, 3 models expect Deepseek runs better than the individual benchmark (which the transaction weight for each model shall be different basing on the portfolio difference). However, taking into the evaluation methodology as we introduced above, we can make more specific and accurate analysis. The result in such salutation is listed in Figure 13.

The performance metrics for the four models—Ernie, Deepseek, Llama, and Spark—reveal significant weaknesses across the board, with all models exhibiting negative annualized yields, high max drawdowns, and deeply negative Sharpe ratios. This suggests that none of these models would be viable for practical trading without substantial improvements.

Ernie, which has the lowest win rate among the group at approximately 48.8 percent, suffers from the worst annualized yield at -97.98 percent and the largest max drawdown at -35.41 percent. Its Sharpe ratio of -28.27 further confirms its poor risk-adjusted performance, meaning it generates extreme losses relative to its volatility. The high volatility (0.000560) relative to the other models exacerbates its instability, making it the riskiest choice despite its slightly better win rate. Deepseek shows a marginally better but still dismal profile, with all data index deeply negative. However, its lower volatility (0.000373) compared to Ernie suggests slightly more stable performance, though this stability does not translate into profitability.

Llama stands out as the least bad performer, with the smallest max drawdown (-9.6 percent), the highest win rate (57.5 percent), and the least negative annualized yield (-59.69 percent). Its Sharpe ratio of -12.70, while still poor, is the least unfavorable among the four, suggesting slightly better risk management. The lowest volatility (0.000290) also indicates more consistent performance, though consistency in losses is hardly a redeeming quality. Rather than that, Spark's metrics with a max drawdown of -28.85 percent, an annualized yield of -95.28 percent, and a Sharpe ratio of -22.37. Even we can observe from the chart that it can outperform the benchmark, its volatility is nearly as high as Ernie's, showing its high risk while with not satisfying outcome.

In summary, none of these models in this scenario demonstrate acceptable performance, with all producing substantial losses and failing to manage risk effectively. Llama is the least problematic, but even its metrics are far from viable. However, adding cost to such high-frequency order may not be accurate, thus we should pay our attention to the next step.

Secondly, we have done the result matrix at the circumstances without the cost. The

Figure 14, Figure 15, Figure 16, Figure 17 are the pnl line separately for four models with the transaction cost not adding into consideration. The performance matrix is as the Figure 18.

Ernie remains problematic across all metrics. The negative Sharpe ratio (-0.72) and -12.62 percent annualized yield reveal a failing strategy, exacerbated by the highest volatility (0.000685) and largest max drawdown (-6.69 percent) among the group. The sub-50 percent win rate (47.80 percent) confirms systematic flaws. Deepseek presents a more conservative profile. With a modest 9.40 percent annualized yield and near-neutral Sharpe ratio (0.99), it avoids extreme losses (max drawdown: -1.57 percent) but also misses substantial gains. The lowest win rate (47.21 percent) suggests its strategy is defensive rather than opportunistic. While stable, its returns may not justify deployment unless capital preservation is the primary objective.

Llama demonstrates better performance, combining high returns with low risk. Additionally, it achieves this with the smallest max drawdown (-1.34 percent) and lowest volatility (0.000294), suggesting a consistently profitable strategy. The win rate which outperform by half and also the largest further confirms its effectiveness. Meanwhile, Spark delivers the highest absolute returns (27.03 percent annualized) but with greater volatility (0.000671) and a larger max drawdown (-3.83 percent). This result may hint those who willing to accept higher risk to make invest decisions for greater potential rewards.

In Conclusion, we can say that in this circumstance Llama is the optimal choice for most scenarios, offering an exceptional balance of returns and risk management. Spark serves as a higher-risk, higher-reward alternative. Deepseek works only for low-risk tolerance situations, while Ernie should be excluded from consideration.

7.2 Daily-level analysis

Rather than the minute level data analysis, we also done the daily frequency analysis. Which is shown in Figure 19. In this data period, Ernie and Spark showed better returns, while Lama and Deepseek exhibit worse. However, four models all did worse rather than Benchmark. While the benchmark gained the outer perform in the zero line, all the models generate least gain rather than the zero line.

Overall, the data underscores the critical role of model efficacy in generating returns, with the conclusion that all the models would require substantial improvements or alternative strategies to avoid losses. The exclusion of transaction costs in this analysis means the actual performance gap could be even wider, as costs would further erode the margins of the already underperforming models.

On the other hand, we can have an deep-in work based on the data evaluation given, to make the numerical proof. The Figure 20, Figure 21, Figure 22, Figure 23 are the pnl line separately for four models with the transaction cost adding into consideration. The performance matrix is as the Figure 24.

Taking cost into consideration, the Figure 25, Figure 21, Figure 27, Figure 28 are the pnl line separately for four models with the transaction cost adding into consideration. The performance matrix is as the Figure 29.

By analyzing both of the performance matrix, we can say it has revealed significant deterioration across all models when transaction costs are introduced, though none of the strategies prove to be valuable in either scenario. Without transaction costs, all four models show negative annualized yields ranging from -19.2 percent (Ernie) to -30.7 percent (Deepseek), with similarly poor Sharpe ratios between -0.82 and -1.56. These weak results worsen substantially with transaction costs, where annualized losses deepen to -30.8 percent (Ernie) to -40.5 percent (Deepseek), and Sharpe ratios plummet further to between -1.51 and -2.28. This consistent pattern gives a solid proof for the further adjustment need, as they fail to generate positive returns even before accounting for real-world trading friction.

Specifically speaking, the impact of transaction costs proves particularly damaging to Deepseek, which experiences the greatest decline in performance. Its Sharpe ratio worsens from -1.56 to -2.28, and annualized yield drops from -30.7 percent to -40.5 percent. While Llama maintains slightly better risk metrics in both scenarios (lower volatility and drawdowns than peers), its Sharpe ratios around -1.49 to -2.21, can remain firmly unacceptable. The models' win rates - all below 40 percent, which even being unchanged by costs - confirm failures in the design itself rather than execution issues.

Notably, Spark shows the most resilience to transaction costs, with its Sharpe ratio declining from -1.00 to -1.68 compared to deeper drawbacks for other models. This relative stability suggests Spark's strategy may involve less turnover. The universal failure of these models, points to potential flaws in their underlying logic - possibly from overfitting, or in poor feature selection. For any of these models to become valuable, substantial structural improvements would be required.

7.3 Final Discusion

The comparative analysis of minute-level and daily-level trading reveals huge differences in model performance, highlighting the critical impact of trading frequency and transaction costs. At the minute level, models like Llama and Spark demonstrate relatively better performance, but when transaction costs are introduced, all models suffer severe backwards. In contrast, daily-level trading shows uniformly poor performance across all models, even without costs, suggesting that the strategies themselves may be unacceptable rather than sensitive to execution frequency.

The minute-level analysis showed that even some models have theoretical promise in a T+0 and costless environment, their real-world applicability drops sharply when transaction costs are considered. This suggests that high-frequency trading demands exceptionally high predictive accuracy to offset costs which models here have least ability. Among the comparison we shall be clear that the information gap do exist as the data awareness is much more sensitive in minute level data as long as the corresponding changes shall be more clear than the daily level, we have failed to extract the real meaningful information beyond the words and sentences. Meanwhile, we can also determine that the daily-level results indicate systemic weaknesses across all models, since none outperform the benchmark even before costs are applied. The findings emphasize that neither minute-level nor daily-level implementations of these models are in value as currently constructed. Future work should explore more specific approaches, blending lower-frequency signals with cost-aware execution, to determine whether these models could achieve sustainable profitability.

8 Further work

8.1 Model choice

Given the shortcomings of the current models, future research should explore alternative architectures and strategies to improve robustness. Potential avenues include specific fine-tuning or RLHF tuned learning-based approaches for adaptive trading, or other hybrid models. Testing these alternatives in both minute-level and daily-level scenarios, with rigorous cost-adjusted evaluations, will be essential to identifying other hidden profitable strategies. Ultimately, the goal should be developing models that not only outperform benchmarks but also remain resilient under real-world trading conditions.

8.2 Strategy choice

The current analysis focuses on a simple quantitative strategy of buying and selling in the next period, which may not fully capture the potential of these models. Future research should explore alternative strategies, such as multi-period holding, or hybrid approaches combining other signals. Additionally, incorporating risk management techniques—such as stop-loss mechanisms or volatility scaling—could improve robustness. The poor performance observed in both minute and daily frequencies suggests that the models may require more sophisticated trading rules to enhance predictive power. Ultimately, since the present results are discouraging, which highlight the need for further refinement in both model design and execution strategy.

8.3 Others

Finally, some other analysis may comes to improve, such as the word choice analysis could be done, by which the need for word selection to be larger and more accurate. Additionally, since the data for news does not consist of the weekend time, further research work may be in need to add these into consideration. Moreover, we can done more experiments to judge the grading policy in the weight calculation, such as re-scoring based on the initial scores, or to implement a single choice rather than just adding up of the score to get the final weight.

9 Self-Reflection

The 5920 project was a valuable experience that strengthened my technical and professional skills. Working on an industry-relevant problem improved my problem-solving and time management abilities, though better initial planning could have reduced stress.

The presentation can help me communicate complex ideas more clearly within a longer time, but more practice would have boosted my confidence. My supervisors provided great guidance, though I should have asked questions sooner to avoid minor delays.

Overall, the project taught me the importance of adaptability, and clear communication. It was a challenging but rewarding experience that prepared me for real-world problem-solving.

References

[1] Yoshua Bengio. Neural net language models. *Scholarpedia*, 3(1):3881, 2008.

[2] Wei Chen, Qiushi Wang, Zefei Long, Xianyin Zhang, Zhongtian Lu, Bingxuan Li, Siyuan Wang, Jiarong Xu, Xiang Bai, Xuanjing Huang, and Zhongyu Wei. Discfinllm: A chinese financial large language model based on multiple experts finetuning. *arXiv preprint arXiv:2310.15205*, 2023.

[3] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[4] Giuseppe Nuti, Mahnoosh Mirghaemi, Philip Treleaven, and Chaiyakorn Yingsaeree. Algorithmic trading. *Computer*, 44(11):61–69, 2011.

[5] OpenAI. Chatgpt. https://chat.openai.com/, 2022.

[6] Kennedy Riaga. *Effects of corporate governance on Stock return of commercial banks listed at Nairobi securities exchange.* PhD thesis, UoN, 2020.

[7] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need.(nips), 2017. *arXiv preprint arXiv:1706.03762*, 10:S0140525X16001837, 2017.

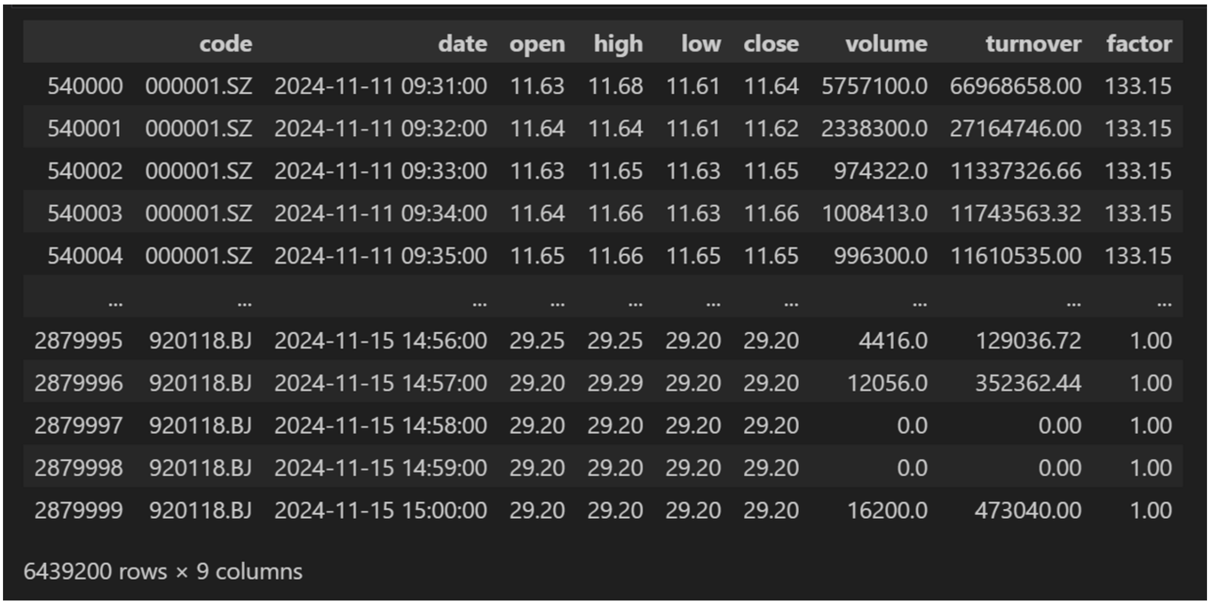
[9] Frank Xing. Designing heterogeneous llm agents for financial sentiment analysis. *ACM Transactions on Management Information Systems*, 2024.

[10] Yi Yang, Mark Christopher Siy UY, and Allen Huang. Finbert: A pretrained language model for financial communications, 2020.

[11] Xinli Yu, Zheng Chen, Yuan Ling, Shujing Dong, Zongyi Liu, and Yanbin Lu. Temporal data meets llm–explainable financial time series forecasting. *arXiv preprint arXiv:2306.11025*, 2023.

[12] Khairul Zharif Zaharudin, Martin R Young, and Wei-Huei Hsu. High-frequency trading: Definition, implications, and controversies. *Journal of Economic Surveys*, 36(1):75–107, 2022.

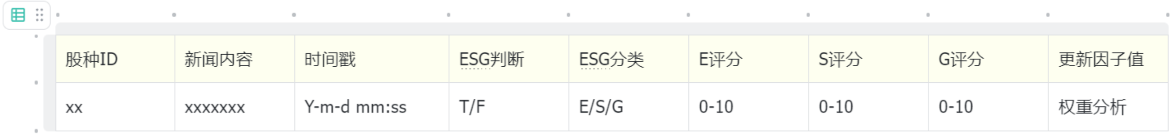
A Data Figure



	code	date	open	high	low	close	volume	turnover	factor
540000	000001.SZ	2024-11-11 09:31:00	11.63	11.68	11.61	11.64	5757100.0	66968658.00	133.15
540001	000001.SZ	2024-11-11 09:32:00	11.64	11.64	11.61	11.62	2338300.0	27164746.00	133.15
540002	000001.SZ	2024-11-11 09:33:00	11.63	11.65	11.63	11.65	974322.0	11337326.66	133.15
540003	000001.SZ	2024-11-11 09:34:00	11.64	11.66	11.63	11.66	1008413.0	11743563.32	133.15
540004	000001.SZ	2024-11-11 09:35:00	11.65	11.66	11.65	11.65	996300.0	11610535.00	133.15
...	...	...	...	...	...	...	...	...	...
2879995	920118.BJ	2024-11-15 14:56:00	29.25	29.25	29.20	29.20	4416.0	129036.72	1.00
2879996	920118.BJ	2024-11-15 14:57:00	29.20	29.29	29.20	29.20	12056.0	352362.44	1.00
2879997	920118.BJ	2024-11-15 14:58:00	29.20	29.20	29.20	29.20	0.0	0.00	1.00
2879998	920118.BJ	2024-11-15 14:59:00	29.20	29.20	29.20	29.20	0.0	0.00	1.00
2879999	920118.BJ	2024-11-15 15:00:00	29.20	29.20	29.20	29.20	16200.0	473040.00	1.00

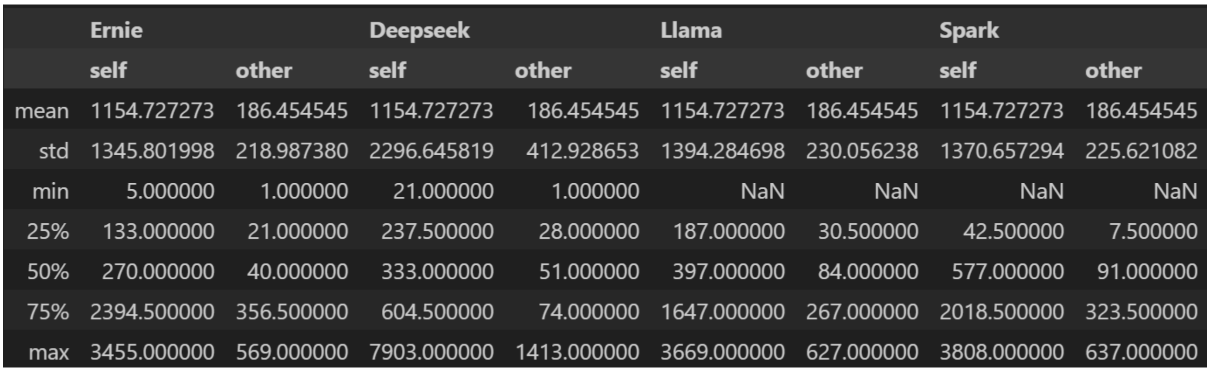
6439200 rows × 9 columns

Figure 5: Minute Level Financial Data



股种ID	新闻内容	时间戳	ESG判断	ESG分类	E评分	S评分	G评分	更新因子值
xx	xxxxxxx	Y-m-d mm:ss	T/F	E/S/G	0-10	0-10	0-10	权重分析

Figure 6: Factor Type



	Ernie		Deepseek		Llama		Spark	
	self	other	self	other	self	other	self	other
mean	1154.727273	186.454545	1154.727273	186.454545	1154.727273	186.454545	1154.727273	186.454545
std	1345.801998	218.987380	2296.645819	412.928653	1394.284698	230.056238	1370.657294	225.621082
min	5.000000	1.000000	21.000000	1.000000	NaN	NaN	NaN	NaN
25%	133.000000	21.000000	237.500000	28.000000	187.000000	30.500000	42.500000	7.500000
50%	270.000000	40.000000	333.000000	51.000000	397.000000	84.000000	577.000000	91.000000
75%	2394.500000	356.500000	604.500000	74.000000	1647.000000	267.000000	2018.500000	323.500000
max	3455.000000	569.000000	7903.000000	1413.000000	3669.000000	627.000000	3808.000000	637.000000

Figure 7: Data Feature

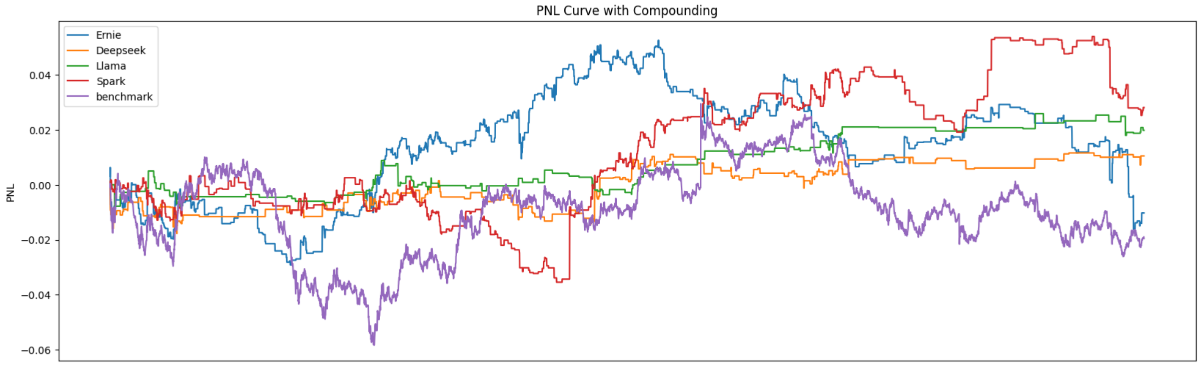


Figure 8: Total pnl without cost-minute

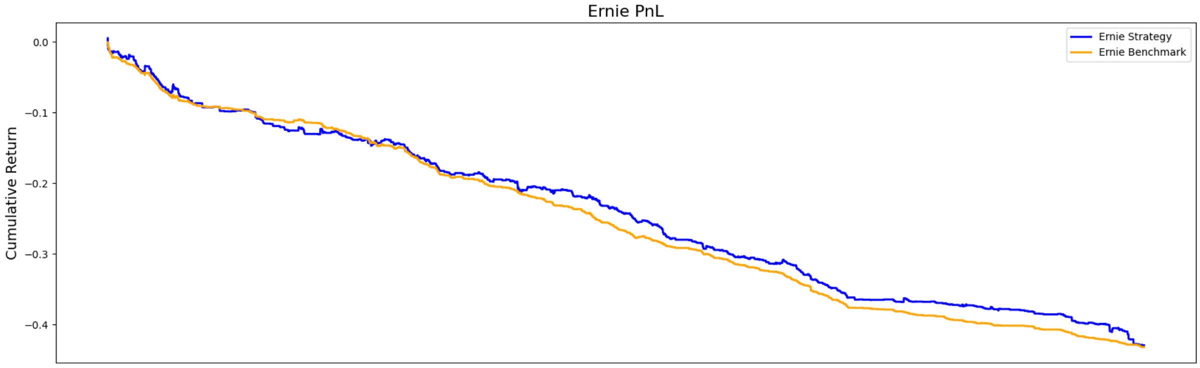


Figure 9: pnl with cost Ernie-minute

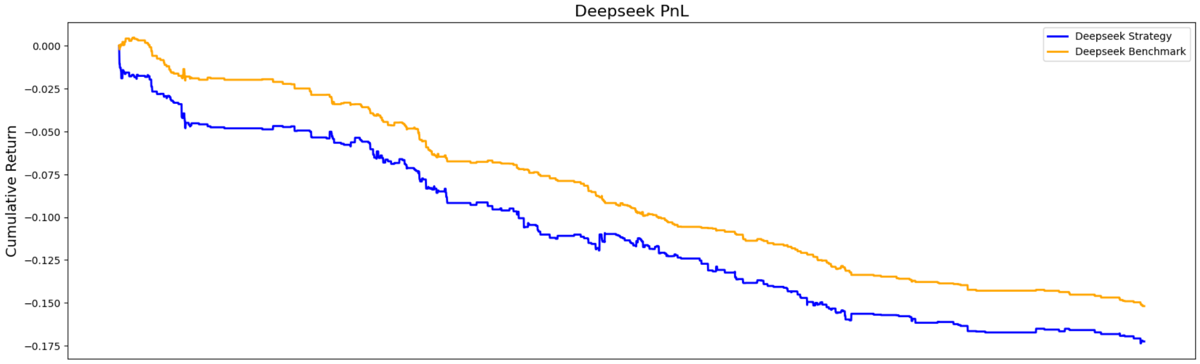


Figure 10: pnl with cost Deepseek-minute

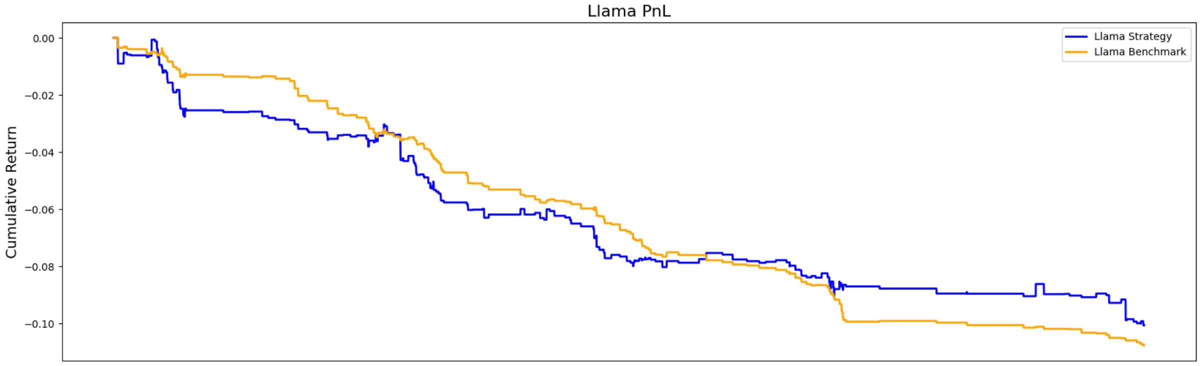


Figure 11: pnl with cost Llama-minute

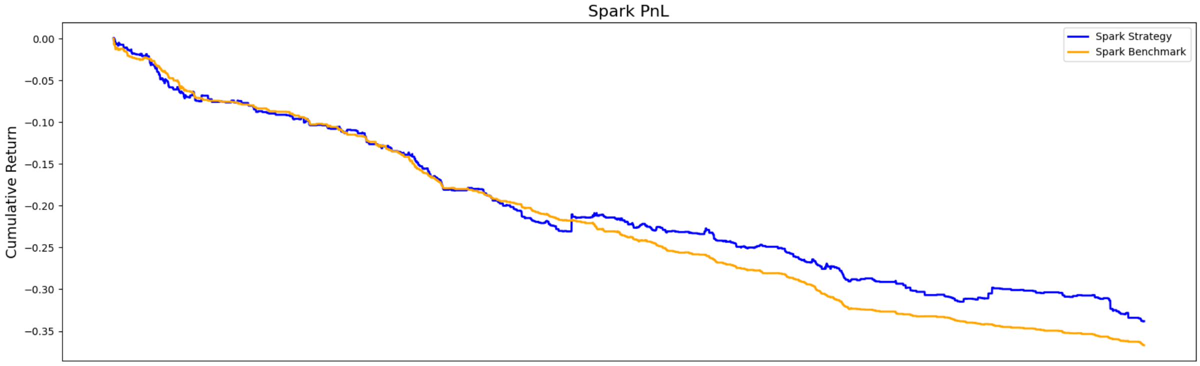
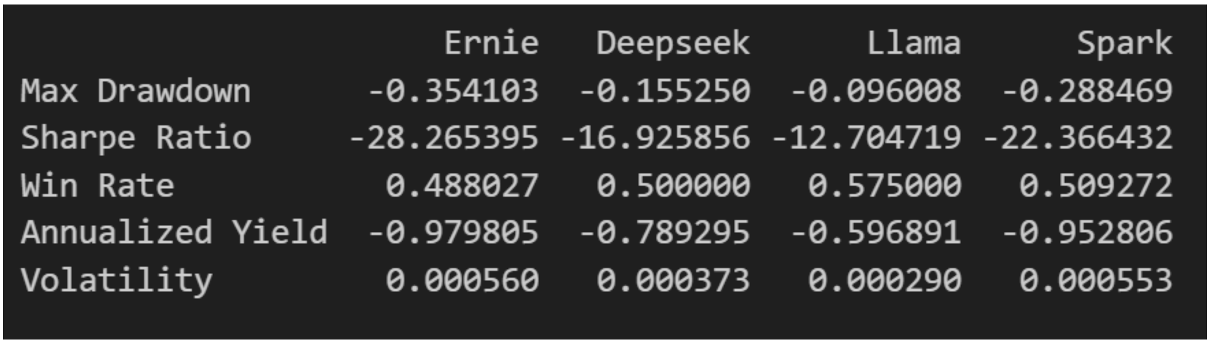


Figure 12: pnl with cost Spark-minute



```
                        Ernie    Deepseek       Llama       Spark
Max Drawdown        -0.354103   -0.155250   -0.096008   -0.288469
Sharpe Ratio       -28.265395  -16.925856  -12.704719  -22.366432
Win Rate             0.488027    0.500000    0.575000    0.509272
Annualized Yield    -0.979805   -0.789295   -0.596891   -0.952806
Volatility           0.000560    0.000373    0.000290    0.000553
```

Figure 13: Data-minute-with

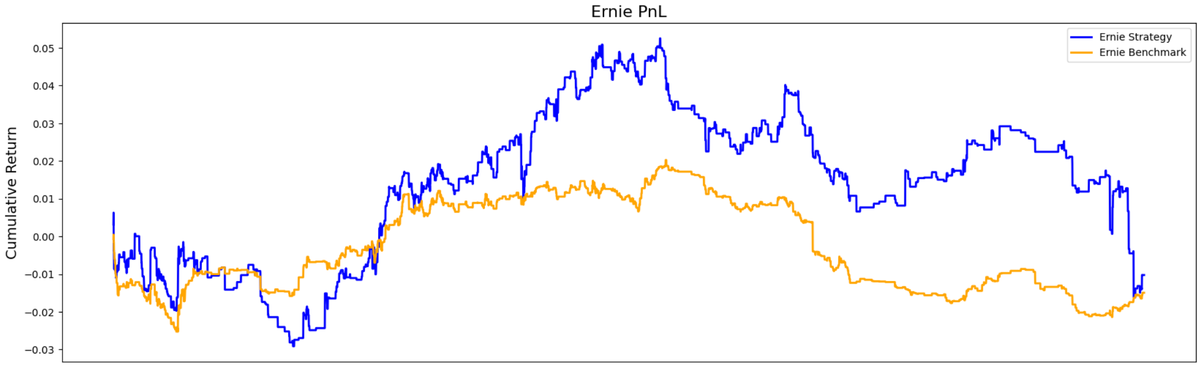


Figure 14: pnl without cost Ernie-minute

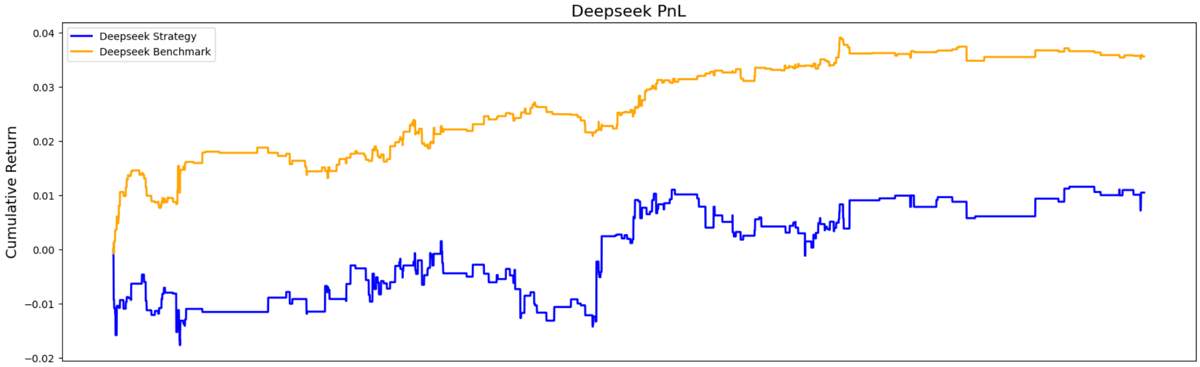


Figure 15: pnl without cost Deepseek-minute

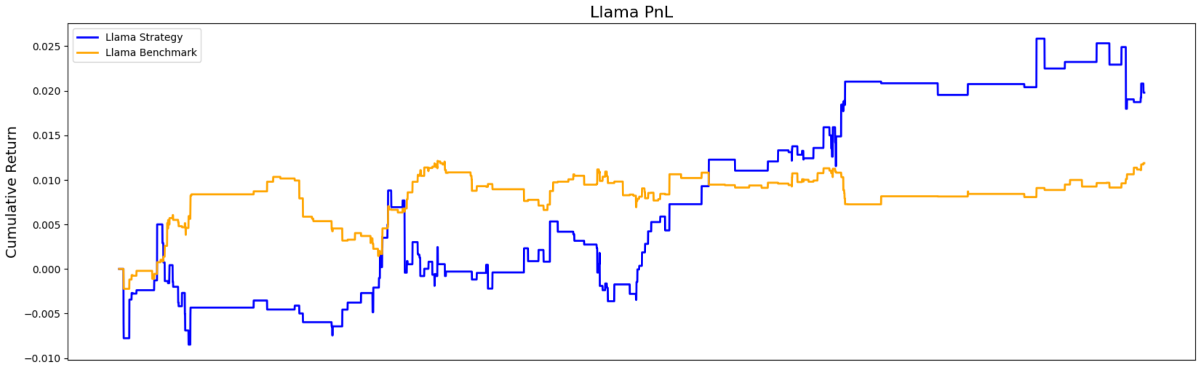


Figure 16: pnl without cost Llama-minute

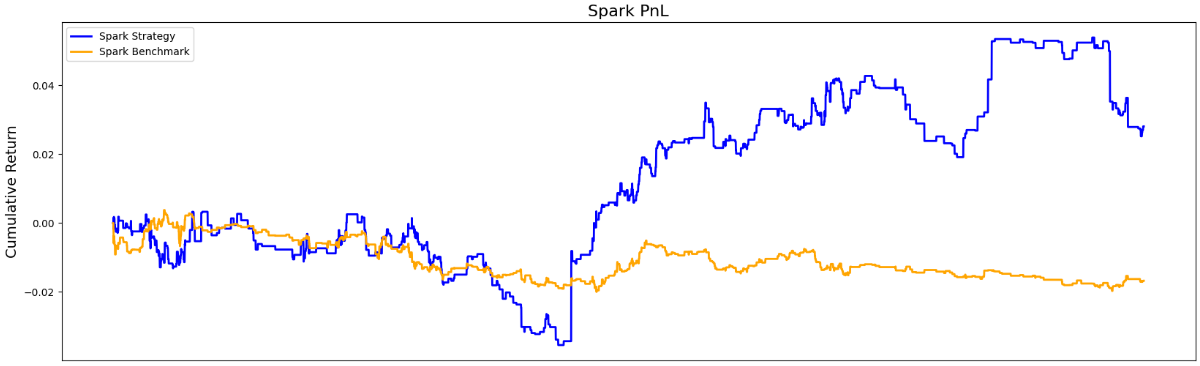
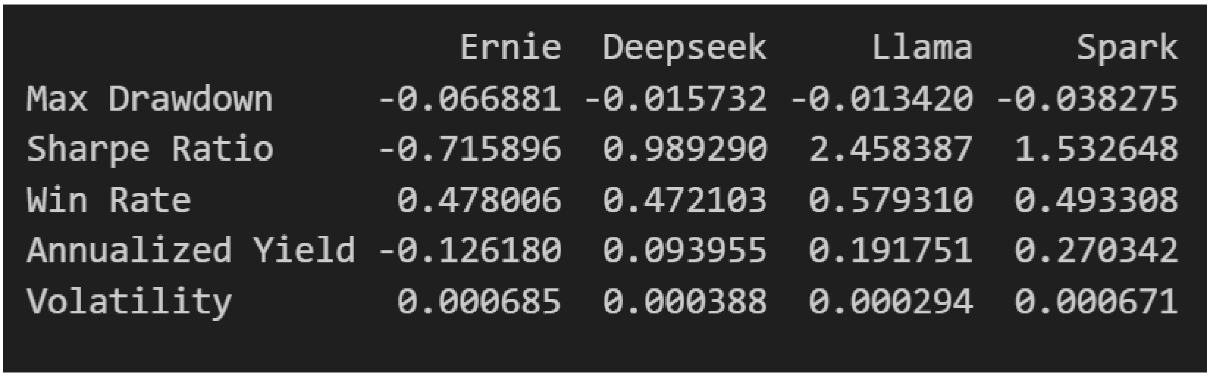


Figure 17: pnl without cost Spark-minute



```
                     Ernie  Deepseek     Llama     Spark
Max Drawdown     -0.066881 -0.015732 -0.013420 -0.038275
Sharpe Ratio     -0.715896  0.989290  2.458387  1.532648
Win Rate          0.478006  0.472103  0.579310  0.493308
Annualized Yield -0.126180  0.093955  0.191751  0.270342
Volatility        0.000685  0.000388  0.000294  0.000671
```

Figure 18: Data-minute-without

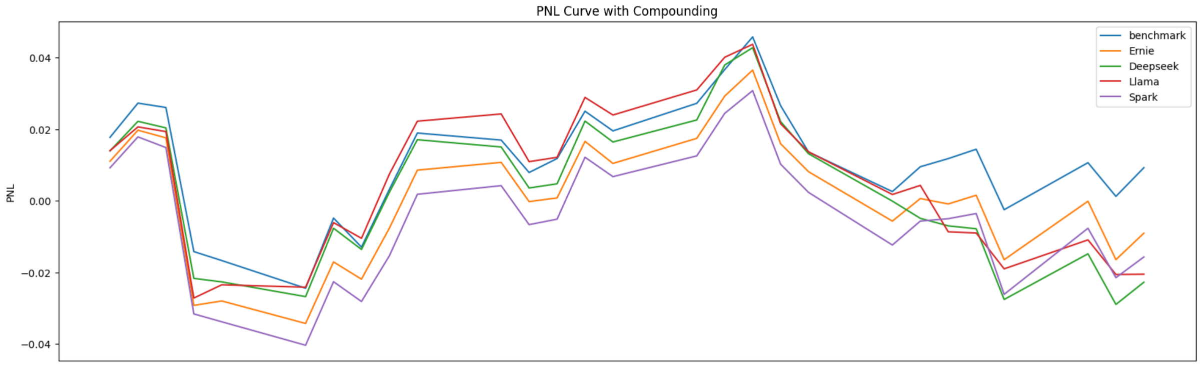


Figure 19: Total pnl without cost-day

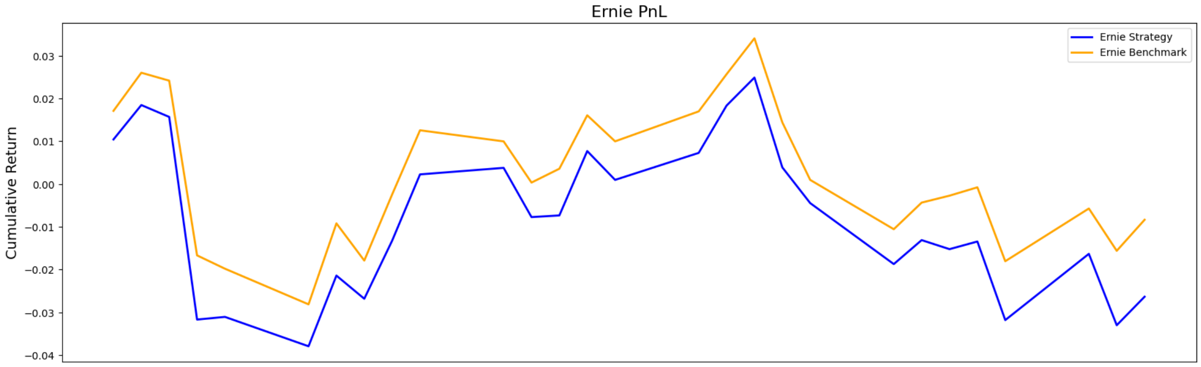


Figure 20: pnl with cost Ernie-day

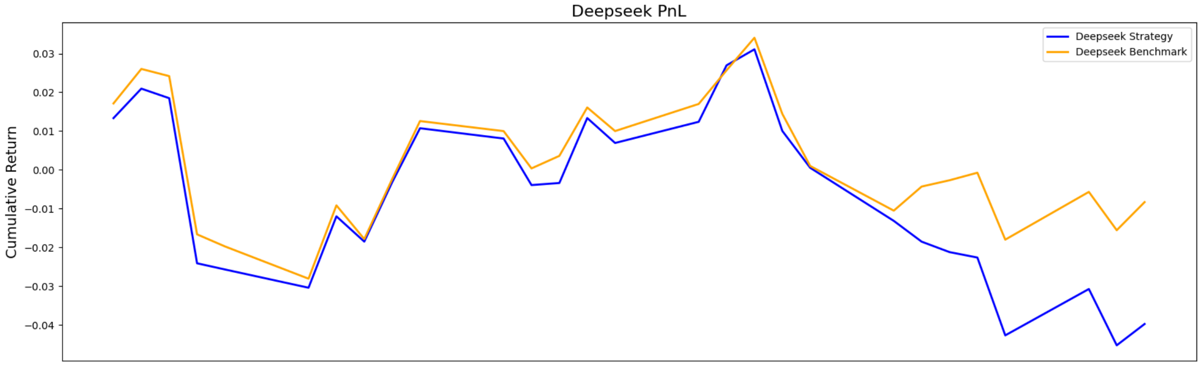


Figure 21: pnl with cost Deepseek-day

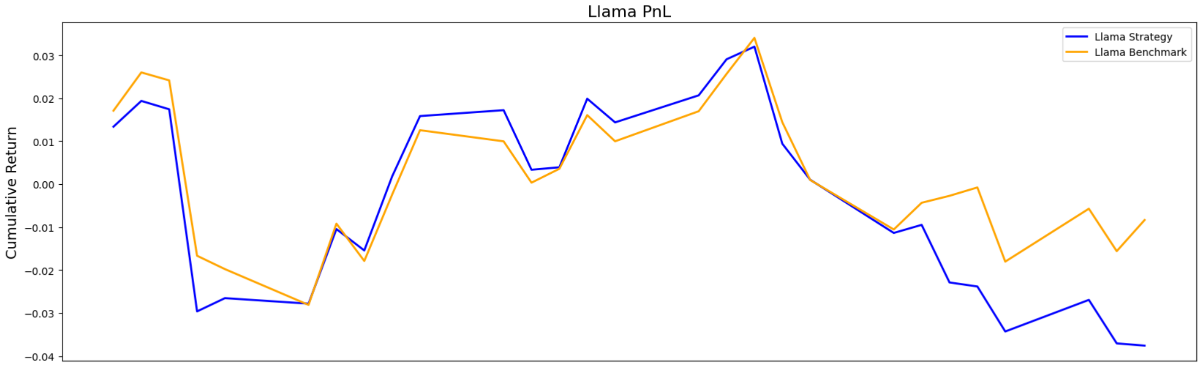


Figure 22: pnl with cost Llama-day

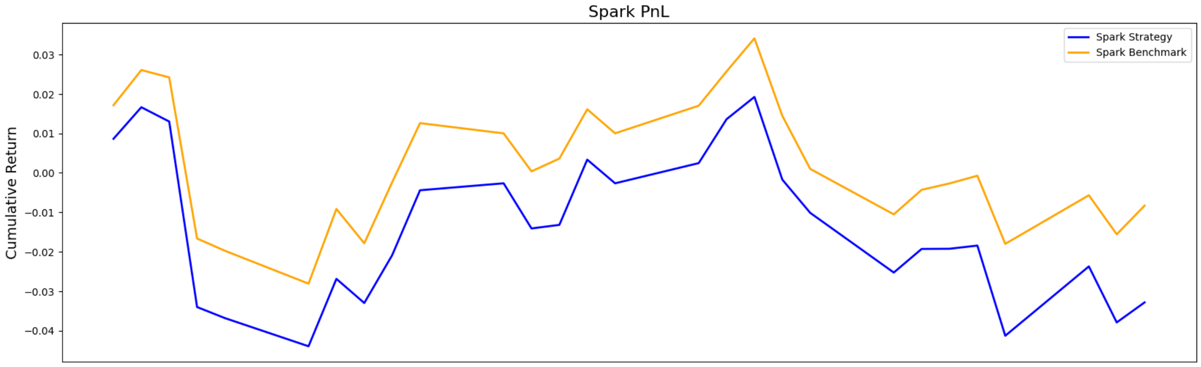
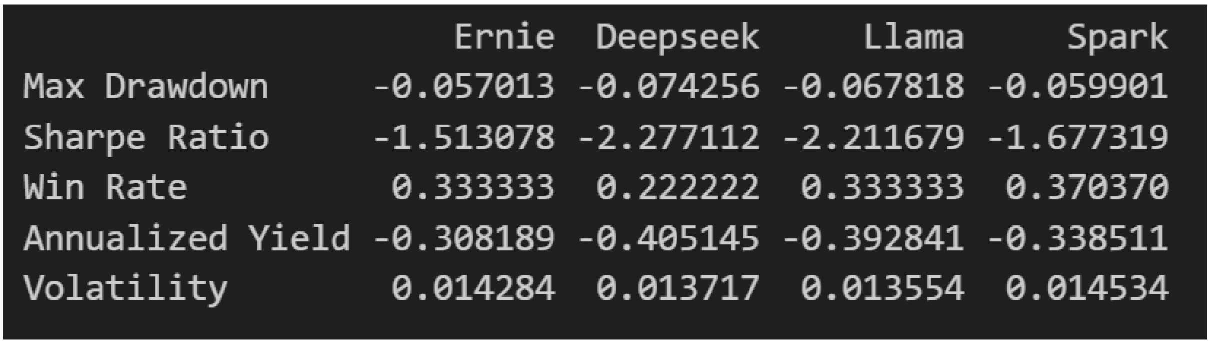


Figure 23: pnl with cost Spark-day



```
                      Ernie  Deepseek     Llama     Spark
Max Drawdown      -0.057013 -0.074256 -0.067818 -0.059901
Sharpe Ratio      -1.513078 -2.277112 -2.211679 -1.677319
Win Rate           0.333333  0.222222  0.333333  0.370370
Annualized Yield -0.308189 -0.405145 -0.392841 -0.338511
Volatility         0.014284  0.013717  0.013554  0.014534
```

Figure 24: Data-day-with

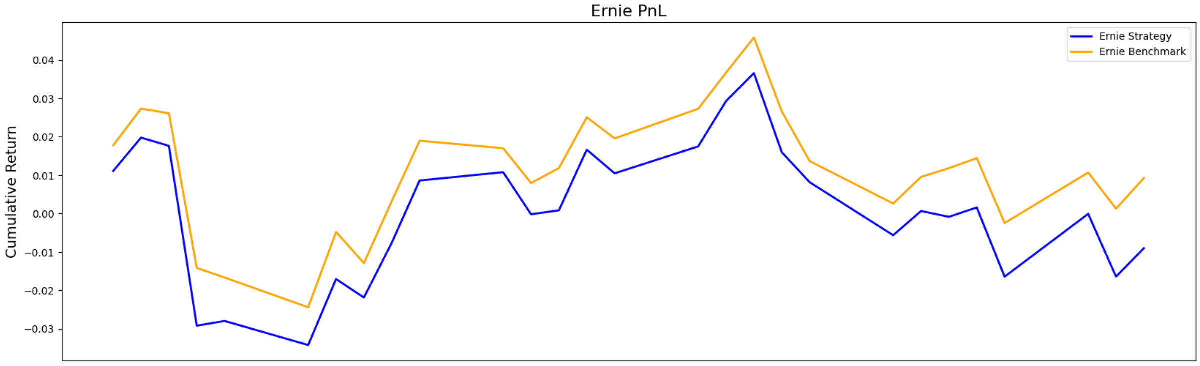


Figure 25: pnl without cost Ernie-day

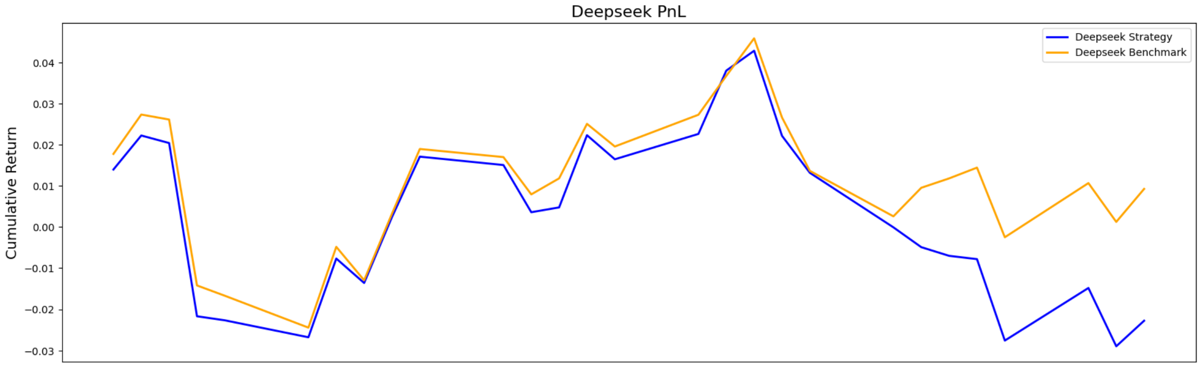


Figure 26: pnl without cost Deepseek-day

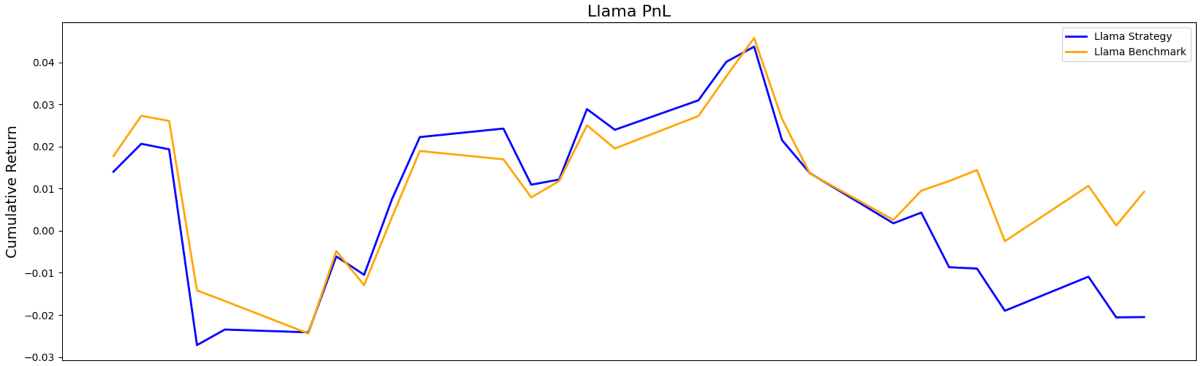


Figure 27: pnl without cost Llama-day

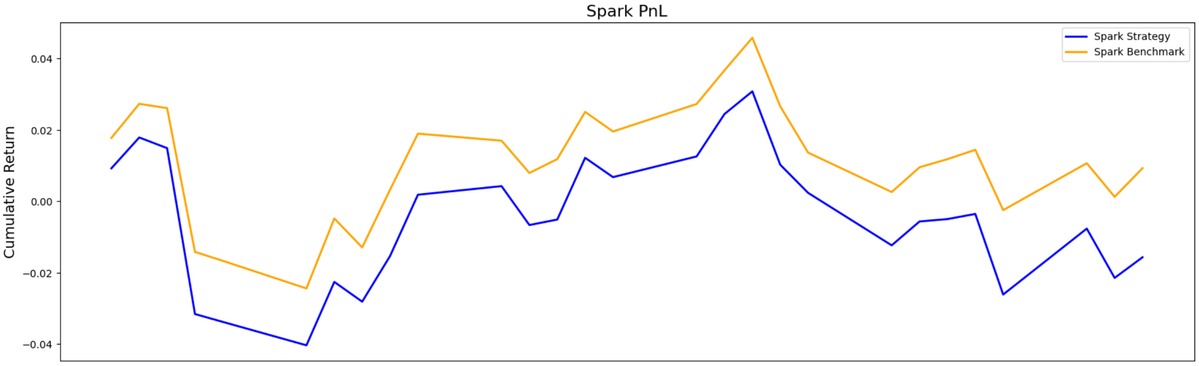
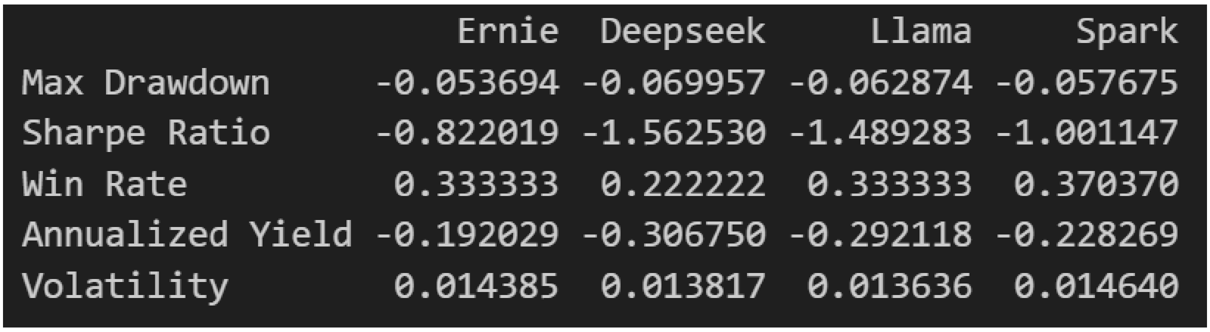


Figure 28: pnl without cost Spark-day



```
                      Ernie  Deepseek     Llama     Spark
Max Drawdown      -0.053694 -0.069957 -0.062874 -0.057675
Sharpe Ratio      -0.822019 -1.562530 -1.489283 -1.001147
Win Rate           0.333333  0.222222  0.333333  0.370370
Annualized Yield  -0.192029 -0.306750 -0.292118 -0.228269
Volatility         0.014385  0.013817  0.013636  0.014640
```

Figure 29: Data-day-without