Assignment 11: LLM Inference for Monetary Policy Classification and Market Analysis

Name: Shiqi Hu

GTID: 904061372

Insights

1. Insights from comparing cumulative returns under different monetary policy stances.

Hawkish Policy Stance Analysis:

- Cumulative Return: The portfolio's cumulative return during hawkish policy stances is approximately 1.30, indicating modest gains during these periods.
- Average Monthly Return: The average monthly return during hawkish periods is 0.67%, translating to an annualized return of 8.39%.
- Observations:
    - The cumulative return and annualized return indicate that the portfolio has a moderate capability to capture gains during hawkish policy periods. This could suggest strategic allocation towards industries and stocks that perform relatively well under tightening monetary conditions.
    - The performance highlights a portfolio that, while still not as robust as during dovish periods, shows resilience and positive gains during hawkish stances.

Dovish Policy Stance Analysis:

- Cumulative Return: The portfolio's cumulative return during dovish policy stances is higher, at approximately 4.29, showing significant outperformance compared to hawkish periods.
- Average Monthly Return: The average monthly return during dovish periods is 1.04%, leading to an annualized return of 13.24%.
- Observations:
    - The consistent outperformance during dovish periods indicates that the portfolio is likely weighted towards growth-oriented or cyclical sectors that thrive in a low-interest-rate environment.
    - This implies an effective capture of economic growth opportunities, benefiting from higher market liquidity and investor risk appetite typical of dovish periods.

Overall Insights:

- Robust Performance: The portfolio's ability to generate positive returns under both hawkish and dovish stances shows its versatility. However, the stronger results during dovish periods suggest an investment strategy that may favor sectors sensitive to growth and lower borrowing costs.
- Outperformance During Dovish Periods: The significant cumulative return and higher annualized return during dovish periods indicate a strategic focus on investments that leverage favorable economic conditions, including industries like Construction and Retail Trade.

2. Reflect on the reliability of using LLM-based classification for financial forecasting.

Strengths:

- Zero-shot and Few-shot Capability: As shown in this FOMC Hawkishness prediction task, LLMs can perform well even with minimal to no task-specific training (zero-shot or few-shot inference). This is particularly useful when labeled data is scarce or difficult to obtain, making it possible to classify financial sentiments or forecast trends based on prompts alone.
- Adaptability: LLMs can understand and classify complex financial language without needing extensive training data, making them adaptable to many financial NLP tasks.

Limitations:

- Limited Access: High-quality LLMs may require significant resources or subscriptions, limiting access for smaller institutions or individual users.
- Domain-specific Limitations: LLMs may not match the accuracy of models trained specifically on financial data, leading to less reliable results in specialized tasks.
- Lack of Transparency: The decision-making process of LLMs can be difficult to interpret, which reduces trust, especially in financial forecasting where explainability is critical.
- Inference Time: The latency of generating outputs with LLMs, especially larger models, can be significant. This can hinder their practical application in real-time financial forecasting where rapid decision-making is crucial.

Conclusion:

While LLM-based classification has the potential to be a powerful tool for financial forecasting, its reliability depends on the context of use. LLMs offer great flexibility with good inferencing performance, especially in the few-shot scenario, but come with limitations related to domain specificity, interpretability, and real-time performance.

Step 0: Background Research

---

1. Define and explain zero-shot inference in the context of language models.

In the context of language models, zero-shot inference is to let the language model generate response according to a task given by the prompt based on its previous knowledge and general understanding of language, without having been explicitly trained on examples of that task in the prompt.

2. Summarize the paper in few sentences in your own words. (Use of ChatGPT not allowed for this part)

This paper aims to evaluate the capabilities of ChatGPT for financial NLP tasks. The authors benchmark the zero-shot performance of ChatGPT-3.5-Turbo, Dolly-V2-12B, and H2O-12B, comparing them to a fine-tuned RoBERTa model across four financial NLP tasks. Their experiments reveal that, while fine-tuned PLMs generally outperform zero-shot ChatGPT on the hawkish-dovish sequence classification task, ChatGPT still demonstrates impressive performance. In financial sentiment analysis, a notable performance gap is observed between fine-tuned PLMs and ChatGPT, especially when the dataset is not publicly available. For financial numerical claim detection, they find that the time required to label a single sample with generative LLMs is significantly higher. Overall, this paper highlights the potential and limitations of ChatGPT, emphasizing the trade-offs between performance, availability of labeled data, and labeling efficiency.

Step 1: Zero-Shot and Few-Shot Inference with Llama-3-70b-Chat Model

---

1. Few-Shot: Experiment with 3-5 example sentences. Analyze the number and type of examples that yield the most accurate results, and explain your selection process.

Selected Few-Shot Examples After Tried Different Sentence Combination:

- Measures of inflation compensation based on TIPS fell in response to the soft reading on core inflation in the November CPI release but subsequently moved up against the backdrop of an improving global growth outlook, higher commodity prices, depreciation of the dollar, and the stronger-than-expected reading on core inflation in the December CPI release. (**1: Hawkish**)
- In their discussion of the balance-of-risks sentence in the press statement to be issued shortly after this meeting, all the members agreed that the latter should continue to express, as it had for every meeting earlier this year, their belief that the risks remained weighted toward rising inflation. (**1: Hawkish**)
- While the underlying demand for residential housing continued to be robust and government outlays evidently were rising, the expansion of consumer spending seemed to have slowed, and outlays for capital spending were still very sluggish in an environment of business uncertainty and pessimism. (**0: Dovish**)
- In their discussion of monetary policy for the period ahead, members agreed that it would be appropriate to maintain the existing highly accommodative stance of monetary policy. (**0: Dovish**)
- Nevertheless, most participants agreed that, although the level of inventories of unsold homes that homebuilders desired was uncertain, the correction of the housing sector was likely to continue to weigh heavily on economic activity through most of this year--somewhat longer than previously expected. (**2: Neutral**)

Explain considerations for selecting examples in few-shot inference for optimal performance.

- According to the zero-shot confusion matrix, it is obvious that the model has poor performance in classifying dovish and hawkish policy states, but performs quite well in classifying neutral state.
- To address this, I decided to select 2 'dovish' and 2 'hawkish' examples, along with 1 'neutral' example.
- These examples span different years to expose the model to varied expressions and structures used in FOMC sentences over time.
- All examples are drawn from a consistent training dataset (lab-manual-mm-train-5768.xlsx) to ensure alignment with the labeling techniques used in the test dataset (lab-manual-mm-test-5768.xlsx).

Do you think your few-shot prompt is optimal? What are some ways you could improve your few-shot prompt? What would the tradeoffs be for that?

- I don't think my few-shot prompt is optimal. Some ways that I think can improve my few-shot prompt are as follows:
    - Increase sample size: Increase the sample size appropriately without violating the principle of "small quantity", especially for dovish and hawkish as these two categories perform poorly. Also increase sample size by selecting different year's expression can help the model have a better understanding about FOMC expressions across the years.
    - Improve sample quality: Choose more challenging or boundary bound sentences, which can help the model better understand the subtle differences between categories.
- Tradeoffs:
    - Complexity vs. Performance: Adding more or complex examples might improve accuracy but could also complicate the model’s processing capabilities, increasing inference time.

- Risk of Overfitting: Tailoring the model too closely to a small set of specific examples might boost its performance on those examples at the cost of reduced generalizability to broader datasets.

2. Parse Model Output:

	sentence	year	label	zero_shot_prediction	few_shot_prediction
0	At the conclusion of the discussion, the Commi...	2009	DOVISH	NEUTRAL	NEUTRAL
1	Moreover, inflation was running at a fairly lo...	2019	DOVISH	DOVISH	DOVISH
2	A few participants judged that while the labor...	2020	NEUTRAL	NEUTRAL	NEUTRAL
3	Inflation was still expected to be somewhat hi...	2020	HAWKISH	NEUTRAL	HAWKISH
4	With the Committee in the process of reviewing...	2007	NEUTRAL	NEUTRAL	NEUTRAL
...	...	...	...	...	...
209	With the risks to the forecast for economic ac...	2003	DOVISH	DOVISH	DOVISH
210	These indicators suggested that the financial ...	2020	NEUTRAL	NEUTRAL	NEUTRAL
211	Job gains had remained solid, and the unemploy...	2005	HAWKISH	NEUTRAL	NEUTRAL
212	Overall inflation was projected to remain subd...	2012	DOVISH	NEUTRAL	NEUTRAL
213	Consumer prices had edged up in recent months,...	2012	DOVISH	NEUTRAL	DOVISH

214 rows × 5 columns

Prediction Performance Evaluation:

- **F1 Score**: The F1 score combines precision and recall into a single metric by taking their harmonic mean. Few-Shot achieves an F1 score of 0.7103, indicating a strong balance between precision and recall. This score represents a noticeable improvement over the zero-shot method (0.5903).
- **Precision**: Precision measures the accuracy of positive predictions. Few-Shot's precision is at 0.7108, showing that when the model predicts a category, it is correct about 71% of the time, which is a substantial increase from the zero-shot 0.6891.
- **Recall**: Recall shows how many actual positives were identified correctly. A 0.7103 recall score of few-shot indicates that the model successfully identifies about 71% of all relevant instances, significantly

better than in the zero-shot method (0.6121).

- **Accuracy**: Accuracy measures the proportion of true results (both true positives and true negatives) among the total number of cases examined. The accuracy rises from 0.6121 in the zero-shot to 0.7103 in the few-shot approach, reflecting an overall improvement in model performance across all classes.
- **Conclusion**: These comparisons clearly illustrate that few-shot learning significantly enhances the model's ability to classify statements more accurately, as evidenced by improvements in all metrics. This improvement likely stems from the model's better understanding of nuances in financial policy statements due to the examples provided during few-shot training.

Step 2: Performance and Latency Analysis

---

1. Results

1.1 Latency Comparison:

- Llama-3-70b-Chat (zero-shot) latency: 1.6953 seconds per sentence (Slowest)
- Llama-3-70b-Chat (few-shot) latency: 1.5703 seconds per sentence (Middle)
- RoBERTa-based PLM (my fine-tuned model from the previous assignment) latency: 1.1764 seconds for 214 sentences, 0.005497 second per sentence (Fastest)

1.2 Performance Comparison:

RoBERTa-based PLM Performance:

- F1 Score: 0.6892 (middle)
- Precision: 0.6985
- Recall: 0.6916
- Accuracy: 0.6916 (middle)
- Confusion Matrix:

	DOVISH	**HAWKISH**	**NEUTRAL**
DOVISH	49	4	16
HAWKISH	4	41	4
NEUTRAL	21	17	58

Few-shot performance:

- F1 Score: 0.7103 (best)
- Precision: 0.7108
- Recall: 0.7103
- Accuracy: 0.7103 (best)
- Confusion Matrix:

	DOVISH	**HAWKISH**	**NEUTRAL**
DOVISH	50	1	18
HAWKISH	3	34	12

	DOVISH	HAWKISH	NEUTRAL
NEUTRAL	17	11	68

Zero-shot performance:

- F1 Score: 0.5903 (lowest)
- Precision: 0.6891
- Recall: 0.6121
- Accuracy: 0.6121 (lowest)
- Confusion Matrix:

	DOVISH	HAWKISH	NEUTRAL
DOVISH	29	0	40
HAWKISH	1	17	31
NEUTRAL	9	2	85

2. Compare the three setups in terms of F1 score, accuracy, and inference time. Document insights about trade-offs between model size, performance, and latency.

Performance Analysis:

- **F1 Score and Accuracy**:
  - The few-shot setup with Llama-3-70b-Chat achieved the highest F1 score (0.7103) and accuracy (0.7103), surpassing even the fine-tuned RoBERTa-based PLM model.
  - Fine-tuned RoBERTa-based PLM achieved an F1 score of 0.6892 and accuracy of 0.6916, placing it between the few-shot Llama setup and the zero-shot Llama setup. While it didn't outperform the few-shot Llama, its performance was more consistent, especially since it was directly fine-tuned on this task.
  - The zero-shot setup of Llama-3-70b-Chat showed the lowest F1 score (0.5903) and accuracy (0.6121). This setup had no examples provided in the prompt, so the model relied purely on its pretrained knowledge, which could explain the lower performance. This emphasizes that while large language models can handle many tasks without examples, a few-shot or fine-tuned approach significantly boosts accuracy.
- **Latency (Inference Time per Sentence)**:
  - RoBERTa-based PLM: With an inference time of 0.005497 seconds per sentence, the RoBERTa-based PLM was the fastest, 286 times faster than Few-shot Llama-3-70b-Chat. Its small size and optimized fine-tuning made it highly efficient, making it suitable for real-time applications where low latency is crucial.
  - Few-shot Llama-3-70b-Chat: The few-shot setup with Llama-3-70b-Chat took around 1.5703 seconds per sentence, making it the middle ground in terms of latency. This setup was more efficient than the zero-shot Llama setup, likely because the few examples provided in the prompt guided the model, helping it achieve faster convergence in processing.
  - Zero-shot Llama-3-70b-Chat: The zero-shot Llama setup had the slowest latency at 1.6953 seconds per sentence, highlighting a significant trade-off between inference time and the ability to handle tasks with minimal guidance. This latency could pose challenges in real-time applications or large-scale deployments, where response speed is essential.

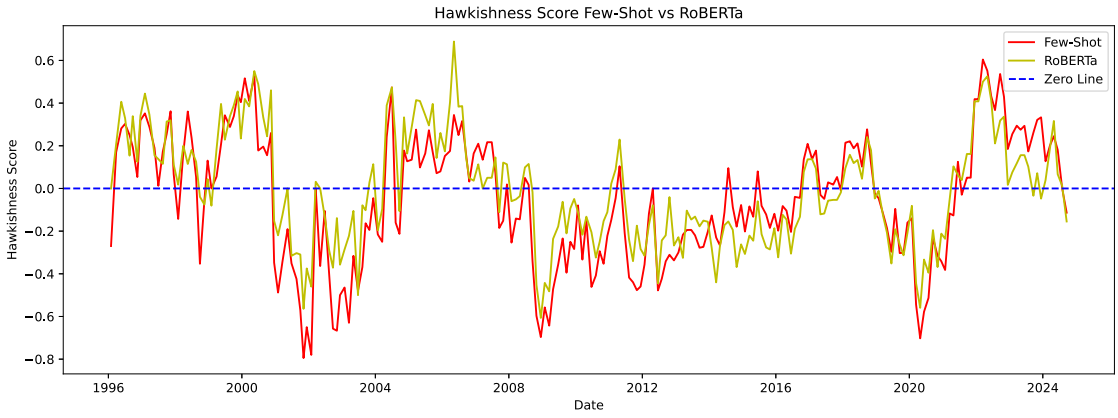
Trade-offs between Model Size, Performance, and Latency

- The few-shot setup for Llama-3-70b-Chat demonstrated that providing just a few examples can significantly enhance model performance without the need for fine-tuning. However, this comes at the cost of longer inference times and increased computational load due to th- model's large size.
- The RoBERTa-based PLM represents a practical balance, achieving decent performance while maintaining extremely low latency. This model is well-suited for scenarios with resource constraints, as it requires less computational power and yields faster results.
- The zero-shot Llama underperformed both F1 score and accuracy compared to the few-shot and fine-tuned approaches. This result highlights that while zero-shot is the most flexible for new categories, it shows poorest performance, significant classification bias and highest latency among the three approaches. Relying solely on zero-shot inference without examples can lead to lower task-specific accuracy.
- Insights:
    - For real-time applications: Use RoBERTa PLM, as its speed advantage is substantial while maintaining acceptable performance.
    - For accuracy-critical tasks: Use Few-shot approach, especially when latency isn't a primary concern.
    - Avoid Zero-shot for this specific task, as it underperforms in both speed and accuracy.

Step 3: Inference on Full Dataset and Constructing Hawkishness Measure

---

Hawkishness Score Comparison: Few-Shot vs RoBERTa



Step 4: Market Analysis and Trading Strategy

---

1. Stock Selection Criteria:

- Historical Performance: Stocks are chosen based on their rolling 12-month return, which captures their performance trend over the past year. This helps in selecting stocks that have demonstrated stability and potential growth, making them suitable for inclusion.
- Sector Classification: Stocks are categorized by sector to align with macroeconomic conditions reflected in hawkishness scores. The strategy targets specific sectors expected to benefit or be harmed under different monetary policies.

- Liquidity and Stability: Only stocks that satisfy the going concern assumption, exhibit a positive price (PRC) are considered. Additionally, stocks must have a market capitalization above a certain threshold to ensure sufficient liquidity and reduce the risk of volatile price movements.

2. Portfolio Composition:

When hawkishness_scores > 0.2 (Hawkish Monetary Policy)

- **Long Positions**: Focus on sectors that historically perform well during hawkish periods, which are:
    - Finance, Insurance, and Real Estate: Beneficiaries of higher interest rates which can boost profit margins.
    - Public Administration: Can remain stable due to government involvement and spending.
- Stock Selection: Choose up to 15 stocks in total across these sectors with the highest 12-month rolling return. These stocks should exhibit resilience or potential in the face of tightening monetary policy.
- **Short Positions**: Short stocks in sectors likely to perform poorly under hawkish policies, which are:
    - Construction: Heavily reliant on borrowing; higher rates increase financing costs.
    - Retail Trade: Consumer spending tends to decline with higher rates, reducing retail revenues.
- Stock Selection: As short stock is associated with more risk, I only short up to 5 stocks in total from these sectors with the lowest 12-month rolling return. These stocks are anticipated to underperform due to increased borrowing costs and reduced consumer spending.

When hawkishness_scores < -0.2 (Dovish Monetary Policy)

- **Long Positions**: Invest in sectors that tend to perform well under dovish policies, which are:
    - Construction: Lower interest rates reduce borrowing costs and boost project funding.
    - Retail Trade: Benefits from increased consumer spending and accessible credit.
- Stock Selection: For each of these sectors, select up to 15 stocks with the highest rolling 12-month return. These stocks are expected to benefit from a dovish monetary policy, which can stimulate consumer spending, borrowing, and investment.
- **Short Positions**: Short stocks in sectors that may not benefit as much from dovish policies, which are:
    - Finance, Insurance, and Real Estate: Potential margin compression due to lower interest rates.
    - Public Administration: Potential mixed effects depending on fiscal responses.
- Stock Selection: Short up to 5 stocks in total from these sectors with the lowest 12-month rolling return. These stocks are less likely to experience significant gains in a dovish environment.

When hawkishness_scores are between -0.2 and 0.2 (Neutral Monetary Policy):

- **Balanced Strategy:**
    - Long Positions: Choose up to 5 stocks from each of the all sectors based on the highest 12-month rolling return. This ensures a diversified approach that captures the best-performing stocks across all sectors.

- Short Positions: Short up to 2 stocks from each of the all sectors with the lowest 12-month rolling return. This balances the strategy by hedging against potential underperformers.

Portfolio Balance:

- The strategy maintains a balanced approach by allocating equal weights to all long and short positions within the portfolio. This ensures that no single stock or sector dominates the portfolio, allowing for diversification and reduced idiosyncratic risk.

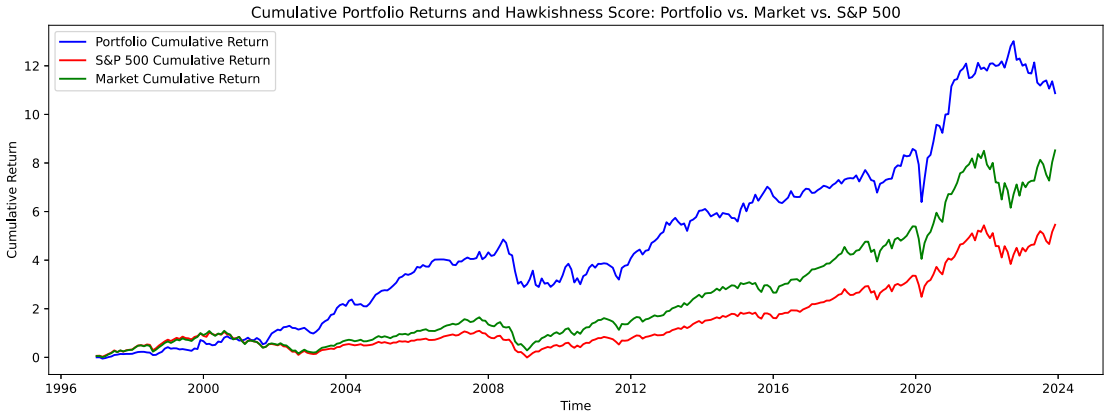
Hawkishness score lag impact is incorporated in Strategy:

- The use of rolling 12-month returns accounts for potential delayed impacts of monetary policy changes, ensuring that stock selection reflects both immediate and lagged market responses. This approach helps the strategy stay adaptable to evolving macroeconomic conditions.

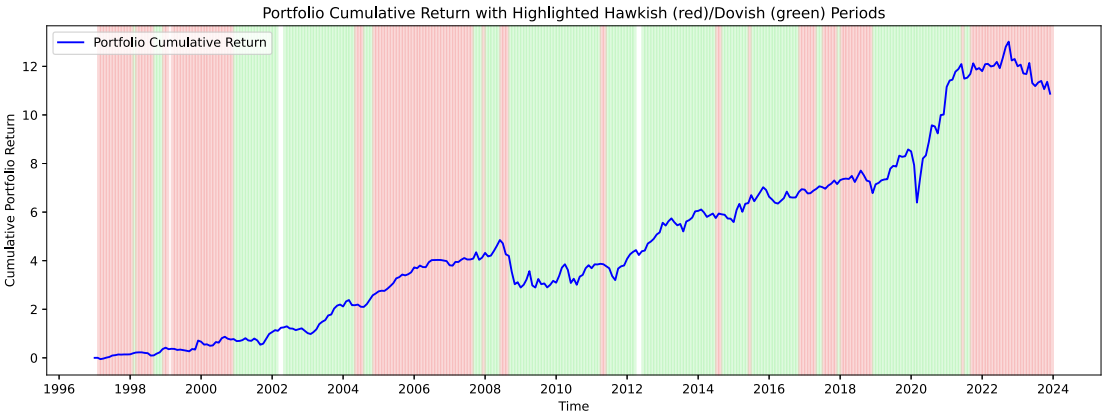
3. Plot it along with cumulative market return for better comparison

Discussion:

- Portfolio Outperformance: The portfolio significantly outpaces both the S&P 500 and the general market, showcasing the effectiveness of its stock selection and sector allocation strategy.
- Comparison Highlights:
  - S&P 500: The portfolio's superior cumulative return indicates better risk-adjusted performance and strategic positioning.
  - Market: The portfolio's targeted approach results in higher returns compared to the broader market, leveraging sector strengths under different monetary policies.
- Resilience: The portfolio's consistent growth underscores its robustness and adaptability in varied market conditions.



4. Comparing cumulative returns under different monetary policy stances



5. Comparison metrics: sharpe ratio, Sortino ratio, max drawdown

- Portfolio Metrics: Sharpe Ratio: 0.6582 | Sortino Ratio: 0.9623 | Max Drawdown: -33.4270%
- S&P 500 Metrics: Sharpe Ratio: 0.4944 | Sortino Ratio: 0.6820 | Max Drawdown: -52.5559%
- Market Metrics from msf data: Sharpe Ratio: 0.5734 | Sortino Ratio: 0.8169 | Max Drawdown: -51.4774%

Discussion:

- Sharpe Ratio:
  - The portfolio's Sharpe Ratio of 0.6582 is notably higher than the S&P 500's 0.4944 and the market's 0.5734 (from msf data). This demonstrates the portfolio's superior risk-adjusted return, showcasing effective risk management and selection strategies that outperform both the broader market and S&P 500.
- Sortino Ratio:
  - With a Sortino Ratio of 0.9623, the portfolio outperforms the S&P 500's 0.6820 and the market's 0.8169. This indicates the portfolio's strong ability to generate returns while minimizing downside risk, appealing to investors who prioritize protection against negative market movements.
- Max Drawdown:
  - The portfolio's maximum drawdown of -33.43% is significantly less severe compared to the S&P 500's -52.56% and the market's -51.48%. This reflects the portfolio's effective risk management strategy in mitigating significant losses, providing better resilience during downturns.