MemAPM: Memory-augmented Large Language Model Agent for Asset Pricing

Anonymous ACL submission

Abstract

In this study, we propose a hybrid asset pricing model, MemAPM, which utilizes a Large Language Model (LLM) agent to refine information from news and augment it with a memory of past refined news. We perform experiments on a two-year span of news and around 70 years of market data, our method outperforms the state-of-the-art machine learning-based asset pricing baselines in multiple portfolio optimization and asset pricing error evaluations. We also performed an ablation study and evaluated the predictive power of the augmented news features for the price movement of individual stocks and economic indicators. The results show the effectiveness of our proposed memory augmentation technique and hybrid asset pricing network architecture.

1 Introduction

The determination of prices on financial assets, such as stocks, has been a focal point in financial economics. It has a significant impact on the economy and society as a whole by improving the allocation of resources. Current asset pricing methods are based on the use of manually designed macroeconomic indicators or company-specific factors as characteristics to forecast future returns (Fama and French, 1992, 2015). Although these methods have been successful in practice in today's market, they have been questioned by the Efficient Market Hypothesis (EMH). According to the EMH, those manual features will gradually lose its predictive power in an efficient market when these predictors are fully discovered and used by traders on the market in the long run.

Due to this rationale, alternative data, such as news, become critical. This is because both the market and society rely heavily on information conveyed through natural language and visual forms. This reliance is also evident in the practical financial world, where discretionary portfolio management continues to have a significant impact on the market (Abis, 2020). In this approach, investment decisions are primarily influenced by the manager's experience and intuition, as they analyze assets and determine their value based on information such as news, investigations, reports, and so on, rather than relying on market data and statistical models.

This phenomenon illustrates two observations. Firstly, linguistic data provide valuable pricing information that is not present in economic factors or market data. Second, despite attempts to integrate semantic analysis and other Natural Language Processing (NLP) techniques into current factor models, these models are unable to completely capture this information. However, utilizing such information is not straightforward as it necessitates financial reasoning and long-term memory of monitoring events and companies' impressions. Moreover, the interplay between news information and manual factors may lead to noise.

In this study, we introduce a novel asset pricing approach, Memory-augment Asset Pricing Models (MemAPM), which combines a Large Language Model (LLM) agent with manual market and economic factors. MemAPM incorporates news features obtained from the LLM agent, which is enhanced by a memory of past refined information, and combines them with manual factors to predict asset returns. To evaluate our approach, we use a dataset comprising two years of news and approximately 70 years of economic and market data. The experimental results show that our approach outperforms existing machine learning-based asset pricing models in terms of the Sharpe ratio of the tangent portfolio, as well as the equal and value weighted long-short portfolio. Furthermore, our method also demonstrates improvement in asset pricing errors for character-section portfolios. The main contributions of this study can be summarized as follows:

- Introducing the use of an LLM agent with long-term memory to augment news input

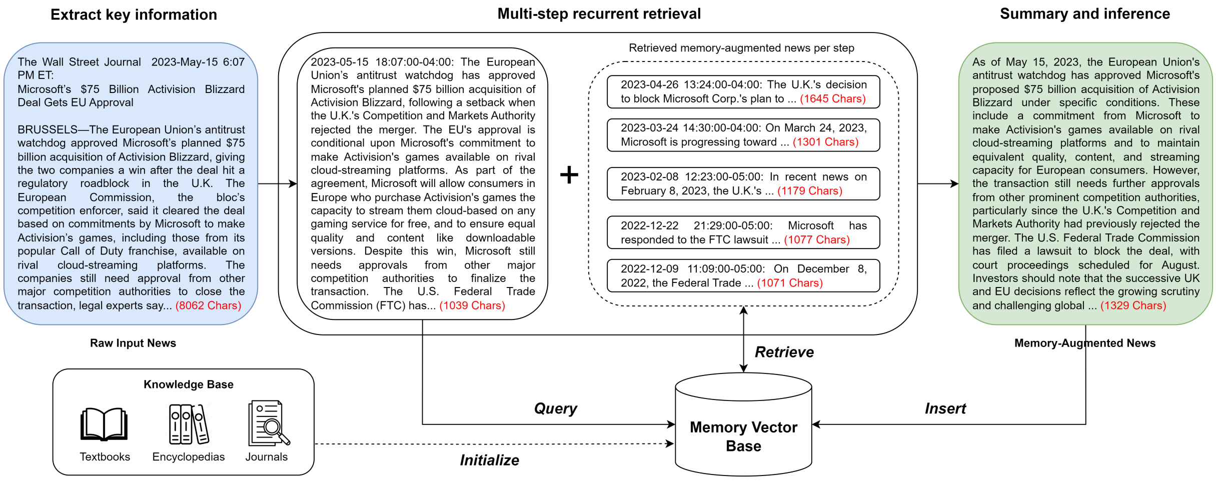

Figure 1: Augment input news by LLM agent with long-term memory.

with past memory and inferences.

- Developing a hybrid asset pricing model that combines the embeddings of augmented news and manual factors to predict returns.
- Conducting experiments and additional data analysis to evaluate performance and gain insights into the proposed method.

The code and data used in our experiments are included in the supplementary material and will be made publicly available after the double-blind review process.

2 Related Work

2.1 Asset Pricing for Security

Asset pricing refers to the field of study that focuses on determining the actual price of financial assets, including securities, which assumes that the market price is wrong. In their groundbreaking work, Sharpe (1964) introduced the Capital Asset Pricing Model (CAPM), which breaks down the expected return of each asset into a linear function of the market return. Over time, various extensions of the CAPM have been proposed. For instance, Merton (1973) incorporated wealth as a state variable to forecast future returns, while Lucas Jr (1978) considered consumption risk as a factor in asset pricing. Building on the idea of decomposing expected return as a linear factor, researchers have also developed multi-factor models. The Fama-French 3-factor (FF3) model, proposed by Fama and French (1992), explains asset returns based on size, leverage, book-to-market equity, and earnings-price ratios. More recently, Fama and French (2015) updated the FF3 model to a 5-factor model. Additionally, Carhart (1997) identified momentum as an additional factor in pricing. Another influential theory in asset pricing is the Arbitrage Pricing Theory (APT) proposed by Ross (1976), which views the actual price of an asset as an equilibrium where arbitrage opportunities do not exist. Furthermore, the Stochastic Discount Factor (SDF) provides a generalized framework for asset pricing, where the price is derived from discounting future cash flows by stochastic factors (Cochrane, 2009).

2.2 Financial Machine Learning

The utilization of machine learning techniques has been introduced to explain the non-linear interactions between various factors and extract valuable information from the extensive "factor zoo" (Feng et al., 2020). Instrumented Principal Component Analysis (IPCA) was developed by Kelly et al. (2020) to estimate latent factors and their loadings from available factor data. In a similar vein, Gu et al. (2020) introduced a deep neural network to predict excess returns of assets. To model latent factors with asset characteristics as covariates, Gu et al. (2021) proposed a conditional autoencoder. Chen et al. (2023) employed GANs (Generative Adversarial Networks) to generate moment conditions for training a model aimed at finding the SDF using the methods of moments. Furthermore, an analysis of the Wall Street Journal Bybee et al. (2021) was conducted to measure the state of the economy. Building upon this analysis, Bybee et al. (2023) further proposed using Latent Direchlet Allocation (LDA) to analyze monthly news topics from The Wall Street Journal as economic factors to determine asset prices.

2.3 Large Language Model Agents

LLM agents are self-governing agents that employ LLMs to process input information and generate output. These agents rely on the emergent capabilities of LLMs, such as reasoning, language comprehension, and function invocation (Achiam et al., 2023). The core of LLM agent programming is the use of prompts, which utilize contextual hinting text to regulate the output of LLMs (Liu et al., 2023). The design of the prompts has a significant impact on task performance. To enhance accuracy, Chain-of-Thoughts (CoT) prompts were developed by Wei et al. (2022) to encourage stepwise reasoning by the agent. Another approach, introduced by Yao et al. (2022), is the ReAct prompting, which allows the agent to refine its output based on the effects of previous attempts. The ReAct framework enables the agent to utilize external tools, such as databases and search engines, for reasoning, thus advancing the development of LLM agents. Memory is another crucial component of LLM agents. Hu et al. (2023) introduced databases as symbolic memories to augment LLM agents. Additionally, Packer et al. (2023) developed an agent capable of storing dialogues in both long- and short-term memory, similar to operating systems. Cheng and Chin (2024) developed an agent that can make "investment" decisions on social science time series based on input news, reports, etc., and knowledge base as well as the Internet.

3 Method

The problem of asset pricing can be represented as the prediction of future returns $r_{t+1,i}$ for a specific asset i, given the current state s_t as hidden factors and the loadings of the asset i on the state A_i, denoted $P(r_{t+1,i}|s_t, A_i)$. In factor-based approaches, the state $s_t \in \mathcal{N}^N$ is a vector consisting of N manually constructed characteristics. These features represent the current state of the market, the economy, or a specific asset. Examples of such features include market excess return, the performance difference between small and large companies, and the difference between high and low book/market companies in the Fama-French 3-factor model (Fama and French, 1992). Alternatively, Bybee et al. (2021) have shown that a collection of business news can be used as an alternative representation of economic status, while (Bybee et al., 2023) utilizes LDA to extract economic predictors for pricing. Building on this concept, we consider the average embedding of news on a given day as the current economic status.

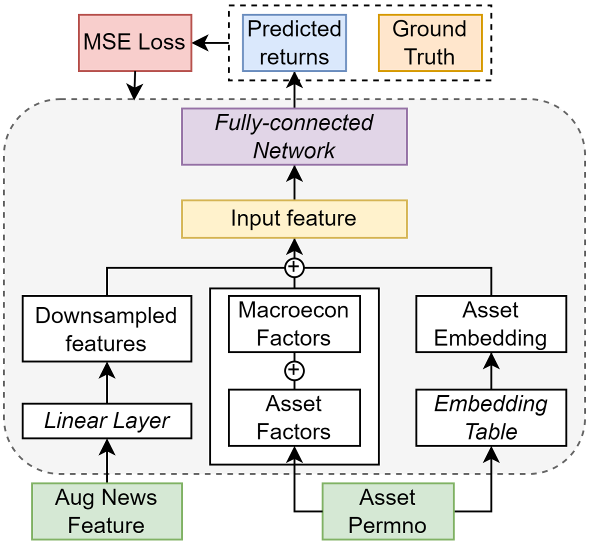

Figure 2: Structure of the prediction network with hybrid predictors of news features and manual factors.

However, professional business news often limits its interpretations and avoids expressing too many opinions in order to reduce bias. This also limits the depth of analysis accompanying the news, leaving ample room for further exploration of the information. It is important to recognize that economic events are often interconnected and should be considered as a chain of events rather than independent news. Interpreting business news involves a reasoning process that is based on background knowledge and related historical events. In this paper, we propose MemAPM, using an LLM agent with long-term memory to enhance input news with reasoning, as discussed in Section 3.1. Subsequently, we combine the augmented input news embeddings with manual factors to introduce our hybrid asset pricing model in Section 3.2.

3.1 Memory Augmentation

At time t, we employ an LLM agent to enhance the input news x_t by transforming it into a memory-augmented news m_t. In our experiments, we utilize GPT-3.5-Turbo (Brown et al., 2020) as the LLM model. The process of memory augmentation is illustrated in Figure 1. Since raw input news is often lengthy, our initial step involves using the LLM to refine the news into a simplified version denoted x'_t that retains only essential information. The prompt we use can be found in Appendix B.

In the next step, we use the simplified news x'_t to retrieve relevant knowledge from memory. The memory is implemented as a vector base and we use ChromaDB [1] for this purpose. We initialize it with the SocioDojo knowledge base (Cheng and Chin, 2024), which comprises textbooks, encyclopedias, and academic journals related to various domains such as economy, finance, business, and other social sciences. The retrieval process is conducted in a multi-step way. In the first step, the simplified news x'_t is used to query the memory, resulting in the retrieval of the top 1 candidate. This candidate is then concatenated with the simplified news, forming a combined news piece. Subsequently, this combined news is used to query the next closest item in the memory excluding the already retrieved item to concatenate on the combined news piece. The same iterative procedure is repeated for a total of K rounds.

After merging the simplified input news with the relevant items in the memory, we employ the LLM to perform a reasoning step on the combined news. This helps us to extract insights that can enhance the information for asset pricing and generate the augmented news m_t. The reasoning prompt used is as follows:

> You are a helpful assistant designed to analyze business news to assist portfolio management. Now, read this latest news and summarize it in one single paragraph, preserving data, datetime of events, and key information, and include new insights for investment using the recommended relevant information: {input}

The augmented news, denoted m_t, will be stored in the memory for future retrieval and will also be utilized in subsequent asset pricing procedures.

3.2 Hybrid Asset Pricing Model

The agent will use a sliding window of size L to produce the state by including the most recent L augmented news $\{m_{t-L+1}, m_{t-L+2}, ..., m_t\}$. It should be emphasized that the time interval between consecutive indices may differ and is not necessarily a fixed time span such as 1 day or 1 minute. This approach allows our method to adaptively adjust the portfolio according to the latest news in real time, as opposed to existing factor models that usually update on a monthly basis, which is consistent with the frequency of updates for most economic indicators.

The state is computed as follows:

$$s_t = \sum_{i=1}^{L} \kappa(t - L + i; t) f_e(m_{t-L+i}) \qquad (1)$$

Where $\kappa(t-L+i;t)$ is an exponential decay factor defined as $\frac{\eta^{\tau_t - \tau_{t-L+i}}}{\sum_{j=1}^{L} \eta^{\tau_t - \tau_{t-L+j}}}$, where τ_t is the date-time corresponds to an index t, in this work, we consider only the date part. The decay coefficient is denoted as $0 < \eta < 1$. The function f_e is an embedding model that generates an embedding vector $e_t \in \mathcal{R}^{d_{emb}}$ on the augmented news. In this work, we utilize the BGE (Xiao et al., 2023) as the embedding model. To construct a hybrid state, we combine the news state with a vector $v_t \in \mathcal{R}^{N_F}$ consisting of N_F economic indicators or company characteristics as manual factors applied to regular factor-based models. The hybrid state is represented as $h_t = [s_t; v_t]$. When introducing asset-specific factors, we can create asset-specific hybrid states $h_{t,i} = [s_t; v_{t,i}]$ for each asset individually. In order to utilize the extensive historical data on manual factors (which covers approximately 70 years), we pre-train the model using the economic factor data prior to 2021 since the news data in our experiment only cover a period of 2 years from 2021 to 2023. We use padding news embeddings (e.g., embedding of "Null") to concatenate with manual factors.

To learn the loading of the asset-specific factor, we introduce the asset embedding $E \in \mathcal{R}^{N_A \times d_{model}}$, where N_A represents the number of assets considered and d_{model} represents the dimension of the embedding. The state is downsampled by $h'_t = \sigma(W_S h_t)$, where σ denotes the ReLU function and $W_S \in \mathcal{R}^{d_{model} \times (d_{emb} + N_F)}$ represents the parameter matrix. The downsampled state is then concatenated with the asset embedding $h_{t,i} = [h'_t; \sigma(E_i)]$, which serves as the hybrid embedding for the asset i. The expected return of the asset i is predicted by $r_{t+1,i} = f_{AP}(h_{t,i})$, where $f_{AP} = f_{AP_i} \circ f_{H_1} ... \circ f_{AP_o}$ represents a multi-layered fully-connected network. Here, $f_{AP_i}(\cdot) = \sigma(W_{AP_i} \cdot)$, with $W_{AP_i} \in \mathcal{R}^{2d_{model} \times d_{model}}$, and $f_{AP_o}(\cdot) = \sigma(W_{AP_o} \cdot)$, with $W_{AP_o} \in \mathcal{R}^{d_{model} \times 1}$. Furthermore, f_{H_k}, with $k \in [1, 2, 3, ...]$, represents hidden layers parameterized by $W_{H_k} \in \mathcal{R}^{\lceil\Updownarrow\wr\sqcap\updownarrow \times \lceil\Updownarrow\wr\sqcap\updownarrow}$. For simplicity, we have omitted batch normalizations, residual connections, and dropout layers. The out-

[1] https://www.trychroma.com/

put of this network represents the predicted expected return for the next time point. In this paper, we predict the return at the end of each trading day and forecast the return for the following trading day. The demonstration of the prediction network is shown in Figure 2.

The loss function that we use is the Mean Square Error (MSE). The training target is formulated as minimizing the average squared difference between the actual return $r_{i,t+1}$ and the predicted return $\hat{r}_{i,t+1}$ for all time-asset pairs, written as

$$\min_{\theta} \frac{1}{N} \sum_{i,t} (r_{i,t+1} - \hat{r}_{i,t+1})^2 \tag{2}$$

This optimization is performed with respect to the collection of all model parameters denoted by θ. The total number of training samples is denoted by N. The ground-truth return of the asset i at time $t+1$ is represented by $\hat{r}_{i,t}$. The model is trained in batches, where B samples are randomly selected and used to update the model in each iteration. The training process consists of a total of T episodes.

4 Experiment

We perform experiments to evaluate the asset pricing performance of the proposed MemAPM. We introduce the setup of the experiment in Section 4.1, then report and analyze the result of the portfolio optimization experiment in Section 4.2, asset pricing error in Section 4.3, we show our ablation study in Section 4.4. We investigate the predictive power of augmented news for the economy in Section 4.5.

4.1 Experiment Setting

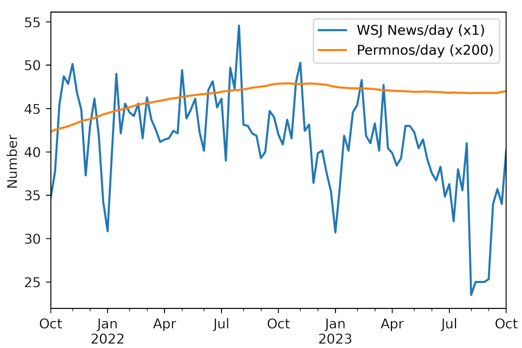

Figure 3: The change in the number of WSJ articles per day after filtering and the number of assets available each day over time.

In our experiment, we utilize a dataset consisting of 2 years' worth of articles from The Wall Street Journal (WSJ), covering the period from September 29, 2021, to September 29, 2023, which beyond the date of latest knowledge in GPT to avoid any potential information leaks. We manually filtered out articles that were not relevant to the business domain, such as those related to travel and lifestyle, based on the category labels provided by WSJ. We obtained the daily return of assets from the Center for Research in Security Prices (CRSP) [2], as well as the daily risk-free return and market return data from Kenneth French's data library [3]. We constructed economic factors and asset-specific factors following Chen et al. (2023). It is worth noting that different factors may have different update frequencies. For factors that were not updated at a given time point, we simply replicated the values from the previous time step. We imputed the missing values using the cross-sectional median following previous methods. We divided the dataset into three subsets: the first 9 months of data were used as a training set, the subsequent 3 months were used as the validation set, and the remaining 1 year was set aside as the test period. Figure 3 provides a visual representation of the changes in the number of articles and assets over time.

We have selected five state-of-the-art asset pricing models as our baseline. NN (Gu et al., 2020) introduced a deep neural network for asset pricing, while IPCA(Kelly et al., 2020) developed instrumental PCA to explore hidden factors and loadings. CA(Gu et al., 2021) proposed a conditional autoencoder, NF (Bybee et al., 2023) utilizes the latent Direchlet allocation of news articles as hidden factors, and CPZ(Chen et al., 2023) applied GAN to solve the stochastic discount factors. We replicated the baselines using the same settings as described in their respective papers and applied the factors they have selected. For both our method and the baselines, we have conducted a hyper-parameter search to compare the best results. We provide the hyper-parameter optimization setting for our method in Appendix A.

4.2 Portfolio Optimization

We first evaluate the Sharpe ratio of the portfolios constructed based on the predicted returns of each asset. Sharpe ratio (SR), denoted $S_p = \frac{\bar{r}_p - \bar{r}_f}{\sigma(r_p)}$, where r_f is the risk-free return, r_p is the portfolio return, and σ denotes the standard deviation. This

[2] https://www.crsp.org/

[3] https://mba.tuck.dartmouth.edu/pages/faculty/ken.french

	SR ↑			MDD (%) ↓		
	TP	EW	VW	TP	EW	VW
NN	3.82	2.83	2.36	4.82	8.12	9.12
IPCA	4.07	2.96	<u>2.66</u>	<u>3.77</u>	<u>5.77</u>	8.63
CA	4.03	2.85	2.55	3.79	6.31	**4.66**
NF	*3.73*	*2.76*	*2.34*	*5.12*	*7.91*	*6.31*
CPZ	<u>4.10</u>	<u>3.02</u>	2.61	4.32	6.27	5.71
Ours	3.86	2.99	2.47	4.36	6.99	6.17
Full	**4.17**	**3.14**	**2.67**	**3.61**	**5.59**	<u>5.05</u>

Table 1: Sharpe Ratio (SR) and Maximal Drawdown (MDD) for Tengency Portfolio (TP), Equal-Weighted (EW) and Value-Weighted (VW) long-short portfolio built based on NN (Gu et al., 2020), IPCA(Kelly et al., 2020), CA(Gu et al., 2021), NF (Bybee et al., 2023), CPZ(Chen et al., 2023), our method without manual factors and the full model with the manual factors. We bold the best results and underlined the second bests.

	avg $\lvert\alpha\rvert$	avg $\lvert t(\alpha)\rvert$	$\frac{\#\lvert t(\alpha)\rvert>1.96}{\#test\ assets}$	GRS
NN	0.83	2.89	0.64	6.89
IPCA	0.76	2.45	0.55	6.38
CA	0.77	2.63	0.52	6.42
NF	*0.89*	*2.77*	*0.62*	*7.32*
CPZ	0.74	**2.44**	0.49	6.77
Ours	0.81	2.64	0.51	6.92
Full	**0.73**	2.49	**0.46**	**6.32**

Table 2: Asset pricing errors for anomaly portfolios with NN (Gu et al., 2020), IPCA(Kelly et al., 2020), CA(Gu et al., 2021), NF (Bybee et al., 2023), CPZ(Chen et al., 2023), our method without manual factors and the full model with the manual factors.

ratio provides a measure of the portfolio's risk-adjusted performance. Additionally, we measure the maximal drawdown, denoted as $MDD(T) = \max_{\tau\in(0,T)}[\max_{t\in(0,\tau)} X(t) - X(\tau)]$, to assess the extreme unexpected scenario of the generated portfolio. The Maximal Drawdown (MDD) is defined as the maximum decline in the total value of the portfolio, $X(t)$, over a given period of time, T. It is calculated as the maximum difference between the peak value of the portfolio, $X(\tau)$, and the lowest value of the portfolio, $X(t)$, within the time interval $(0, \tau)$.

We examine three common methods for constructing a portfolio. The first method is the tangency portfolio, which determines the weight vector as $w_t = E_t[R^e_{t+1} {R^e_{t+1}}^T]^{-1} E_t[R^e_{t+1}]$. In this equation, R^e_{t+1} represents the predicted excess returns of each asset. The tangency portfolio represents the optimal portfolio in a frictionless trading environment. The second and third methods are based on the long-short decile portfolio. In these methods, the assets are ranked by their expected returns. The second method involves longing the top 10% assets and shorting the bottom 10% assets. In the third method, the assets are equally longed or shorted. Alternatively, assets can be weighted according to their market capitalization. These methods aim to create portfolios that are more applicable in real trading environments.

The experiment results for the SR and MDD can be found in Table 1. Our full model, which incorporates both news and manual factors, achieved state-of-the-art SR in all three portfolio construction methods, while also achieving the best or second best MDD. On the other hand, our method that solely relies on news as input, without considering manual factors, showed a significant improvement in the SR compared to a similar language-based asset pricing baseline NF. Specifically, the improvement was 3.5%, 8.3%, and 5.6% across the three portfolio construction methods. Moreover, the maximal drawdown exhibited enhancements of 14.8%, 11.6%, and 2.2%. It is worth noting that NF utilized WSJ news spanning 33 years, while we only used 2 years of news data.

4.3 Asset Pricing Error

We conducted further analysis on the asset pricing performance of the proposed method. In line with the study by Bybee et al. (2023), we selected 78 anomaly portfolios as test assets. These portfolios were constructed using 78 characteristics, including standard anomaly characteristics such as idiosyncratic volatility, accruals, short-term reversal, and others, as identified by Gu et al. (2020). To evaluate performance, we calculated several metrics. One of them is the average absolute alpha, denoted as $\hat{\alpha_i}$. This metric was computed by dividing the expected value of the estimated error term, $\hat{\epsilon_{t,i}}$, by the square root of the average squared returns, $E[R_{t,i}]$, for all quantile-sorted portfolios. This normalization was done to account for the differences in average returns between portfolios. Additionally, we computed the t-value for the results and examined the ratio of t-values greater than 1.96. These measures provide insight into the statistical significance of the findings. Furthermore, we conducted a Gibbons, Ross, and Shanken (GRS) Test (Gibbons et al., 1989) to determine if the regression intercepts, represented by $\alpha_1, \alpha_2, ..., \alpha_n$, are jointly zero. This test helps to assess the overall significance of the intercepts in the regression analysis.

Table 2 presents the results. Our approach, which exclusively utilizes news input, shows a 9.0% increase in alpha and a 4.7% enhancement in t-value compared to NF. Furthermore, there is a 17.7% decrease in the ratio of pricing result with a t-value greater than 1.96, and a 5.5% improvement in the GRS test. When combined with manual factors, our model achieves state-of-the-art alpha, GRS, and high t-value ratio while maintaining a similar t-value as the state-of-the-art methods.

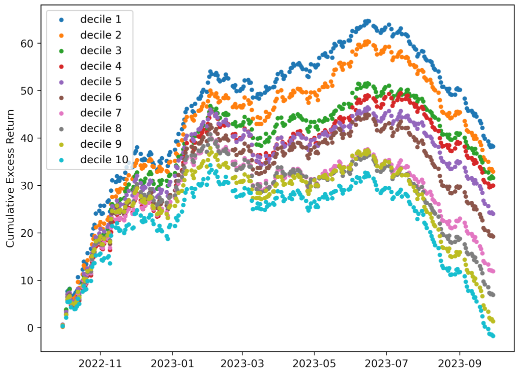

Figure 4: Cumulative excess return for decile portfolios.

We proceed to evaluate the suggested approach by applying it to the pricing of decile portfolios. This involves sorting the assets based on their predicted returns and then creating portfolios for each decile. The cumulative excess return over time is depicted in Figure 4. The figure clearly illustrates that each decile forms a distinct ranking of returns, indicating that the proposed method accurately predicts returns across different levels.

4.4 Ablation Study

In our ablation study, we examine the effects of incorporating LLM features, implementing memory augmentation, and utilizing the asset embedding trick. We compare these approaches with NF, which is a news-based asset pricing baseline, as well as an NF model that includes manual factors, similar to our complete model. Table 3 presents the results of our analysis. When we consider a base model without any augmentation, we observe only marginal enhancements compared to NF which de-

	SR	MDD	avg $\lvert\alpha\rvert$	avg $\lvert t(\alpha)\rvert$
NF	*2.76*	*7.91*	*0.89*	*2.77*
+ Factors	*2.66*	*8.82*	*0.97*	*2.86*
Base	2.82	6.03	0.88	2.72
+Mem	2.94	5.89	0.83	2.66
+Emb	2.88	6.42	0.86	2.71
Ours	2.99	6.99	0.81	2.64
w/o PT	3.03	7.12	0.79	2.62
Full	**3.14**	**5.59**	**0.73**	**2.49**

Table 3: Ablation study of MemAPM and comparison with NF (Bybee et al., 2023). "+Factors" means concatenate the NF news features with the same manual factors in our full model. "Base" is our model without memory augmentation, asset embedding, and manual factors, "+Mem" and "+Emb" is the base model with memory and asset embedding, respectively. "w/o PT" is our model with manual factors, but without pre-training with historical data. We use equal-weighted long-short portfolios for all models.

notes the gain from the improvement of embedding model. However, when we incorporate memory augmentation, we observe a 4.3% improvement in SR. Similarly, the utilization of asset embedding leads to a 2.1% improvement. When both memory augmentation and asset embedding are employed together, we achieve a 6.0% improvement.

Our approach also improves the representation of the non-linear relationships between manual factors and news features. The inclusion of manual factors negatively affects the performance of NF, which aligns with the findings of Bybee et al. (2023) where the introduction of Fama-French factors reduces the Sharpe ratio. However, our method leverages the benefits of manual factors even without pretraining, demonstrating its ability to effectively filter out potential noise and enhance the information extracted from the combined features. Furthermore, we observe that pretraining enhances the performance by 2.8%.

4.5 News as Economic Predictor

We evaluate the ability of the augmented news features to predict the daily percentage change in macroeconomic indicators obtained from the FRED database[4]. These indicators include the stock market (SP500), the market yield on U.S. Treasury Securities at 10-Year constant maturity

[4]https://fred.stlouisfed.org/

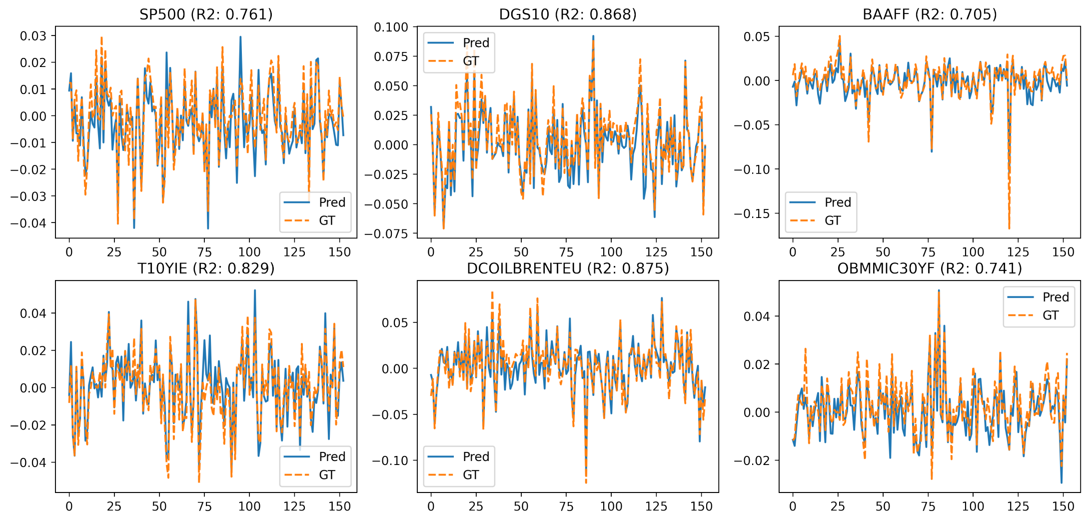

Figure 5: Use news features as the predictor to predict daily percentage change of the economic indicators.

(DGS10), Moody's seasoned Baa corporate bond minus federal funds rate (BAAFF), the 10-year breakeven inflation rate (T10YIE), Brent crude oil prices (DCOILBRENTEU), and the 30-year fixed-rate conforming mortgage index (OBMMIC30YF). The results are presented in Figure 5. The predicted results demonstrate a high level of accuracy, as indicated by the high R2 score. This suggests that the news contains valuable information to predict economic indicators. We visualized the keywords of the news associated with positive and negative expected returns in Appendix C. In Appendix D, we provide an additional analysis of the predictive power of augmented news for individual stock price movements.

5 Discussion

In order to further improve the proposed method, one possible improvement is to enhance the agent initialization process. Currently, the SocioDojo knowledge base contains general knowledge in social science presented as plain text chunks. However, in reality, humans do not process information in this manner. Humans assimilate knowledge from text and transform it into a more organized and structured format, rather than simply memorizing isolated text fragments. Additionally, certain texts may contain formulas or examples that should be further refined and used in specific ways.

In addition to initialization, there is room for improvement in the retrieval process. Currently, the model only retrieves based on unconditional distance. However, finding a relevant business event may require considering multiple factors, such as economic conditions. By implementing a conditional retrieval method, the performance can be enhanced. Furthermore, efficiency should also be taken into account, as the amount of relevant information may be substantial. In the current method, we retrieve less than 10 snippets, which is partly due to the challenge of condensing news into a more concise form.

Finally, the ability to connect to the Internet and access a wider range of information sources, including those found in SocioDojo, can be a valuable approach. This is because the public information available in the real market extends beyond just news sources. Additionally, it is important to take into account multimodal information such as diagrams and figures.

6 Conclusion

In this study, we proposed MemAPM, which utilized an LLM agent with long-term memory to augment input business news. We combined news embeddings with manual factors to create a hybrid feature for asset pricing. MemAPM outperformed state-of-the-art methods in various assessments, including portfolio optimization and asset pricing error. We also conducted a thorough analysis to examine the contributions of each component, as well as a study on the predictive ability of news on economic indicators. We are confident that our research can advance the understanding of the interplay between language information and manual economic and company factors, ultimately leading to greater efficiency in the economy.

Limitations

Our experiments only focus on the US market and English news, which may potentially impact model performance in lower-resources languages. In order to exclude the information leak, we only applied news data after September 2021, thus the study is restricted to the 2 years period after this time, so it is unknown how well the proposed method can be generalized beyond this period, although we use a large test split where half of the dataset was applied as the test set. Finally, public information in the stock market includes not only news, but also reports, reports from social networks, academic journals, opinions from experts, etc.; we do not cover those information as discussed in Section 5.

Ethics Statement

We do not identify any ethical concerns in our approach. Our study does not involve any human participation. Furthermore, the application area of our method is not directly linked to humans, reducing the risk of abuse or misuse. In fact, considering a wider range of information, our method has the potential to enhance market efficiency, resulting in economic benefits for society.

References

Simona Abis. 2020. Man vs. machine: Quantitative and discretionary equity management. *Machine: Quantitative and Discretionary Equity Management (October 23, 2020)*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Leland Bybee, Bryan Kelly, and Yinan Su. 2023. Narrative asset pricing: Interpretable systematic risk factors from news text. *The Review of Financial Studies*, 36(12):4759–4787.

Leland Bybee, Bryan T Kelly, Asaf Manela, and Dacheng Xiu. 2021. Business news and business cycles. Technical report, National Bureau of Economic Research.

Mark M Carhart. 1997. On persistence in mutual fund performance. *The Journal of finance*, 52(1):57–82.

Luyang Chen, Markus Pelger, and Jason Zhu. 2023. Deep learning in asset pricing. *Management Science*.

Junyan Cheng and Peter Chin. 2024. Sociodojo: Building lifelong analytical agents with real-world text and time series. In *The Twelfth International Conference on Learning Representations*.

John Cochrane. 2009. *Asset pricing: Revised edition*. Princeton university press.

William Falcon and The PyTorch Lightning team. 2019. PyTorch Lightning.

Eugene F Fama and Kenneth R French. 1992. The cross-section of expected stock returns. *the Journal of Finance*, 47(2):427–465.

Eugene F Fama and Kenneth R French. 2015. A five-factor asset pricing model. *Journal of financial economics*, 116(1):1–22.

Guanhao Feng, Stefano Giglio, and Dacheng Xiu. 2020. Taming the factor zoo: A test of new factors. *The Journal of Finance*, 75(3):1327–1370.

Michael R Gibbons, Stephen A Ross, and Jay Shanken. 1989. A test of the efficiency of a given portfolio. *Econometrica: Journal of the Econometric Society*, pages 1121–1152.

Shihao Gu, Bryan Kelly, and Dacheng Xiu. 2020. Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273.

Shihao Gu, Bryan Kelly, and Dacheng Xiu. 2021. Autoencoder asset pricing models. *Journal of Econometrics*, 222(1):429–450.

Chenxu Hu, Jie Fu, Chenzhuang Du, Simian Luo, Junbo Zhao, and Hang Zhao. 2023. Chatdb: Augmenting llms with databases as their symbolic memory. *arXiv preprint arXiv:2306.03901*.

Bryan Kelly, Dacheng Xiu, et al. 2023. Financial machine learning. *Foundations and Trends® in Finance*, 13(3-4):205–363.

Bryan T Kelly, Seth Pruitt, and Yinan Su. 2020. Instrumented principal component analysis. *Available at SSRN 2983919*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Robert E Lucas Jr. 1978. Asset prices in an exchange economy. *Econometrica: journal of the Econometric Society*, pages 1429–1445.

Robert C Merton. 1973. An intertemporal capital asset pricing model. *Econometrica: Journal of the Econometric Society*, pages 867–887.

Charles Packer, Vivian Fang, Shishir G Patil, Kevin Lin, Sarah Wooders, and Joseph E Gonzalez. 2023. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*.

Stephen A Ross. 1976. The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 13(3):341–360.

William F Sharpe. 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, 19(3):425–442.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

A Hyperparam Search

Parameter	Distribution
Learning rate	{1e-3,1e-4,5e-4,5e-3}
d_{model}	{128,256,512,768,1024}
d_{emb}	{128,256,512,768,1024}
Epochs	{50,100,150,200}
Hidden Layers	{0,1,2,3,4,5}
Dropout rate	$U(0, 0.3)$
Batch size	$U_{log}(32, 1024, 8)$
η	$U(0.9, 1)$
L	{100,200,300,500}

Table 4: Distributions for the key hyperparameters in the hyperparameter search.

To conduct hyperparameter searches for our approach, we utilize Weights & Biases Sweep (Biewald, 2020). The distribution of the empirically significant parameters in our hyperparameter search is presented in Table 4. The notation $U(a, b)$ denotes a uniform distribution ranging from a to b, while $U_{log}(a, b, r)$ represents a logarithmic uniform distribution with a base of r between a and b. The evaluation criteria for our method is based on the Sharpe ratio of the equal-weight long-short portfolio.

We conducted our experiments on our internal clusters, and the major workload has the following configuration:

- 2 × Intel Xeon Silver 4410Y Processor with 12-Core 2.0GHz 30 MB Cache
- 512GB 4800MHz DDR5 RAM
- 2 × NVIDIA L40 Ada GPUs (no NVLink)

We employed PyTorch Lightning (Falcon and The PyTorch Lightning team, 2019) for parallel training.

B Prompt for simplifying news

This simplification of the news input discussed in Section 3.1 is achieved through the following prompt:

> You are a helpful assistant designed to analyze business news. You need to use brief language to describe key information and preserve key data in the news. Now, analyze the following news: {input}

C Dataset visualization

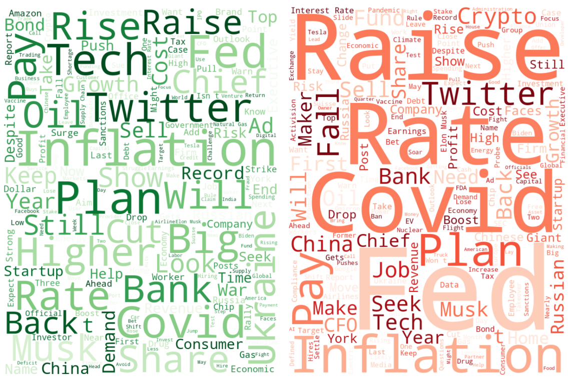

Figure 6: Visualization of the key words found in the titles of news articles on the days when the predicted return is positive (left) and when the predicted return is negative (right).

We initiated our analysis by examining the main topics covered in the news articles in our dataset over time. These topics were identified based on

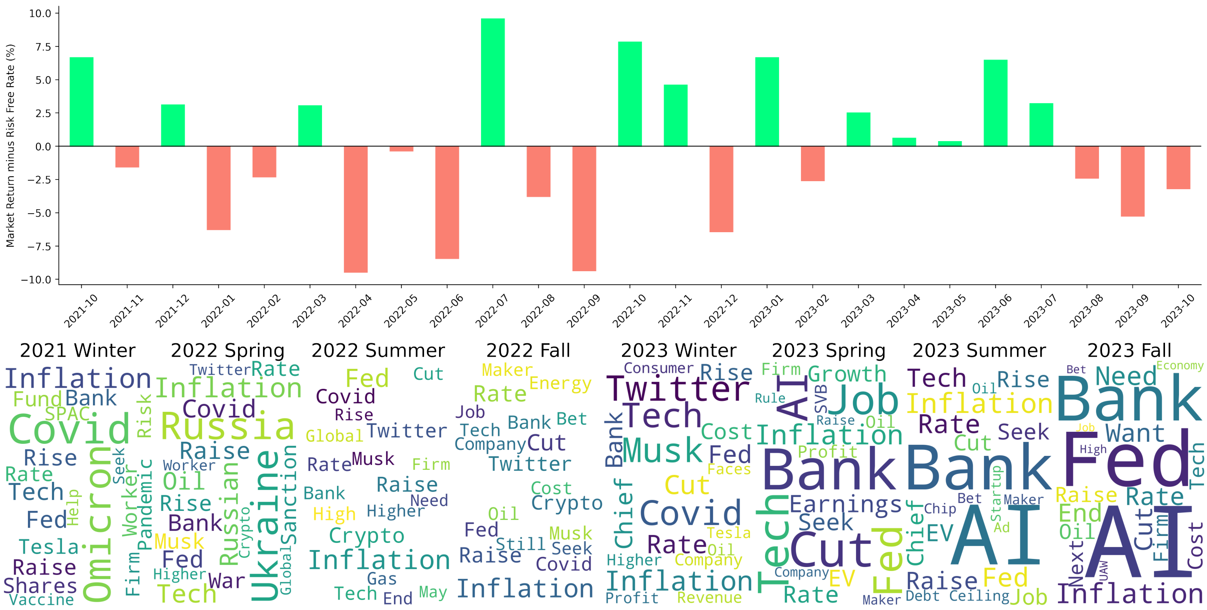

Figure 7: The word cloud of topics of the Wall Street Journal business news changing over time in our dataset (Bottom) compared to the corresponding risk adjusted market return over time (Top).

the titles of the news articles for each season. We eliminated common words such as "US," "Stock," and "Market" as they did not effectively represent the event's topic. The resulting word cloud is presented in Figure 7. It is evident that the economy is primarily influenced by various events, with a gradual decrease in the emphasis on COVID. Instead, the focus shifted towards controlling inflation and the decisions made by the FED. The banking crisis at the beginning of 2023 then became the new focal point, followed by the recognition of AI as a driving force for the economy, primarily due to the success of LLMs. It shows that these event trends have the potential to serve as robust predictors for economic indicators and the market. It is also reflected in Section 4.5 where we assessed that the news articles have great predictive capability for economic indicators.

In Figure 6, we visually represent the keywords present in the titles of news articles on days with negative and positive predicted returns. The visualization aligns well with human intuition, as events such as the FED rate hike, COVID, and concerns about inflation have had the largest negative impact on the market in the past two years, while factors such as technology, Twitter, and efforts to control inflation have contributed to market growth.

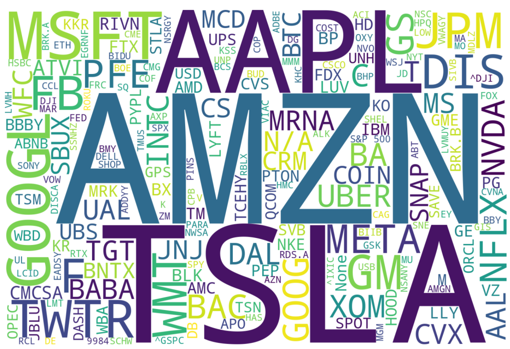

Figure 8: The most frequent mentioned stock tickers in the news.

D News as Stock Price Predictor

We further evaluate the predictive power of the augmented news features for the price movement of individual stocks. We use GPT to analyze the relevant tickers for each news using the prompt below:

> You are a helpful assistant designed to analyze the business news. You need to extract the stock tickers of the companies most closely related to the news. If there is no relevant ticker, return an empty list. You should never make up a ticker that does not exist. Now, analyze the following news: {input}

The tickers associated with the news in our

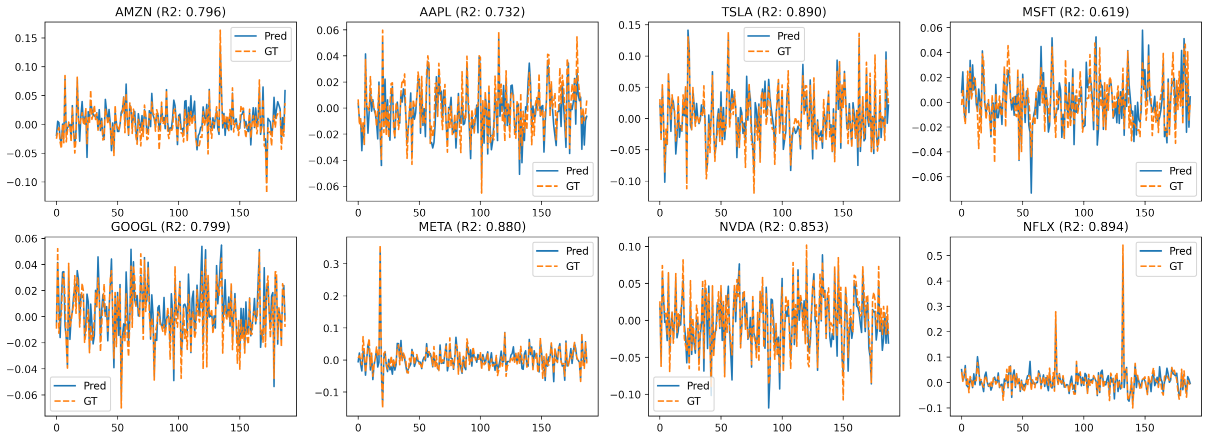

Figure 9: Predict the price movement of stocks in focus using augmented news features as predictors.

dataset are shown in Figure 8. It is clear that technology stocks have been the main focus of the market in the past two years, which is in line with our expectations and the strong performance of these stocks. We have chosen 8 stocks of interest and used augmented news features as predictors to forecast their daily percentage price change. The results are presented in Figure 9, with the corresponding R2 scores indicated in brackets. We observe high accuracy and R2 scores for all the selected stocks, suggesting that news can play a crucial role in predicting stock prices. However, it is important to note that due to the non-stationarity and the risk of overfitting, stock price prediction cannot replace asset pricing (Kelly et al., 2023), but it can provide insights into the predictive power of news in stock market.