Large Language Models and Return Prediction in China

First Version: November 7th, 2023
This Version: April 5th, 2024

Lin Tan, Huihang Wu and Xiaoyan Zhang[*]

Abstract

We examine whether large language models (LLMs) can extract contextualized representation of Chinese news articles and predict stock returns. The LLMs we examine include BERT, RoBERTa, FinBERT, Baichuan, ChatGLM and their ensemble model. We find that tones and return forecasts extracted by LLMs from news significantly predict future returns. The equal- and value-weighted long minus short portfolios yield annualized returns of 90% and 69% on average for the ensemble model. Given that these news articles are public information, the predictive power lasts about two days. More interestingly, the signals extracted by LLMs contain information about firm fundamentals, and can predict the aggressiveness of future trades. The predictive power is noticeably stronger for firms with less efficient information environment, such as firms with lower market cap, shorting volume, institutional and state ownership. These results suggest that LLMs are helpful in capturing under-processed information in public news, for firms with less efficient information environment, and thus contribute to overall market efficiency.

Keywords: return prediction, news articles, large language models, information efficiency, Chinese stock market
JEL Codes: C52, C55, C58, G0, G1, G17

[*] Lin Tan (tanl.19@pbcsf.tsinghua.edu.cn), Huihang Wu (wuhh@pbcsf.tsinghua.edu.cn), and Xiaoyan Zhang (zhangxiaoyan@pbcsf.tsinghua.edu.cn) are all at the PBC School of Finance at Tsinghua University. We thank Lin William Cong, Byoung-Hyoun Hwang, Xing Liu, Xiumin Martin, Dacheng Xiu, Nianhang Xu, Guofu Zhou, and seminar participants at Tsinghua University for their helpful comments. We appreciate Zicheng Wang's research assistance. We gratefully acknowledge financial support from the National Natural Science Foundation of China [Grant No. 72350710220]. All remaining errors are our own. Corresponding author: Xiaoyan Zhang, PBC School of Finance, 43 Chengfu Road, Beijing, China, 100083, zhangxiaoyan@pbcsf.tsinghua.edu.cn.

Large Language Models and Return Prediction in China

Abstract

We examine whether large language models (LLMs) can extract contextualized representation of Chinese news articles and predict stock returns. The LLMs we examine include BERT, RoBERTa, FinBERT, Baichuan, ChatGLM and their ensemble model. We find that tones and return forecasts extracted by LLMs from news significantly predict future returns. The equal- and value-weighted long minus short portfolios yield annualized returns of 90% and 69% on average for the ensemble model. Given that these news articles are public information, the predictive power lasts about two days. More interestingly, the signals extracted by LLMs contain information about firm fundamentals, and can predict the aggressiveness of future trades. The predictive power is noticeably stronger for firms with less efficient information environment, such as firms with lower market cap, shorting volume, institutional and state ownership. These results suggest that LLMs are helpful in capturing under-processed information in public news, for firms with less efficient information environment, and thus contribute to overall market efficiency.

Keywords: return prediction, news articles, large language models, information efficiency, Chinese stock market
JEL Codes: C52, C55, C58, G0, G1, G17

1. Introduction

Text data contain rich information for firm valuation, asset pricing, and investment decisions (Tetlock, 2007; Tetlock et al., 2008; Loughran and McDonald, 2011). Efficiently extracting valuable signals from high-dimensional and unstructured text data is a daunting empirical challenge. Unlike tabular numerical data, text data from news articles consist of semantics, word order, and cross-word relations that are not easily measurable. Large language models (LLMs) pre-trained on massive text corpora using deep neural networks offer state-of-the-art capabilities to address this challenge. Recent work, such as Chen, Kelly, and Xiu (2023), demonstrates the power of LLMs in extracting sentiment and forecasting returns for US and international stocks.

In this study, we examine the capabilities of LLMs in extracting Chinese news texts and forecasting stock returns. Applying LLMs to the Chinese stock market is important, yet challenging. First, Chinese language is used by billions of people, and contains a lot of new information about the economy. Second, Chinese, as a language, is materially different from English and might pose a bigger challenge for signal extraction. A major difficulty of processing Chinese text lies in handling the ambiguities of word segmentation and accurately identifying word boundaries based on the surrounding linguistic context.[1] Due to the language barrier, most previous studies, which rely largely on simplistic natural language processing (NLP) methods, focus on English and other languages rather than Chinese. The powerful LLMs currently provide an exciting opportunity for efficiently processing various languages, while their application for

[1] For instance, depending on the context, the phrase “兵兵球拍卖完了” can be segmented as “兵兵球/拍卖/完了” (ping pong auction over) or “兵兵/球拍/卖/完了” (ping pong paddle sold out).

Chinese capital markets remains unexamined.

In this study, we directly examine whether LLMs can extract signals from Chinese, and how these signals affect information processing and price discovery. In particular, we focus on two research questions. First, do news tones and text features estimated using LLMs predict stock returns in China? Second, how fast do news assimilate prices and what is the implication for market efficiency?

Since this study is based on Chinese news in China, here we provide a few important pieces of background information about Chinese capital market, which significantly differ from those of developed markets. First, retail trading is much more prevalent in China, accounting for 80% of daily trading volumes, according to Jones et al. (2023). In contrast, institutions are much more important in developed countries. Meanwhile, both Song (2020) and Titman et al. (2022) point out that Chinese population mostly has low financial literacy. Given hundreds of millions of these retail investors in the capital market, it might not be surprising to find these retail investors have difficulty in processing public information and trading on public news, as documented in Jones et al. (2023), which make it interesting to understand whether LLMs can help to improve investors' performances. Second, as pointed out by Carpenter et al. (2021), even though the overall market information efficiency level is still low in China, it has been gradually improving. It is exciting to understand whether and how LLMs can potentially affect the overall information efficiency.

Our text dataset is the ChinaScope SmarTag, which contains 28 million news articles between January 2008 and December 2023. The news sources include financial medias, government websites, and WeChat official accounts. We choose this dataset because it has the widest coverage of news sources and provides full content of every news article. We consider five LLMs, which are BERT, FinBERT, RoBERTa, Baichuan, and ChatGLM. These five models are selected based on two considerations: representativeness and performance. For instance, BERT-type models (BERT, FinBERT, RoBERTa) underpins modern NLP and pioneers the transformer architecture with hundreds of thousands of parameters (Devlin et al., 2018). For the best performing models among the latest open-source LLMs, ChatGLM (Zeng et al., 2023) and Baichuan (Yang et al., 2023) models are Chinese LLMs with strong published results, open-sourced architectures and weights, accessible documentation, and wide usage. In addition, to obtain representative predictor based on the five individual LLMs, we construct an ensemble model that aggregates their signals.

To answer the first research questions of return prediction, we proceed in three steps. First, we use pre-trained LLMs to convert news text information into numerical vector representations. Second, we use these vector representations to form news tone and construct return forecasts. Third, we use the news tone and return forecasts to directly predict stock returns by forming long-short portfolios based on news tone and return forecasts. The annualized value-weighted returns for the long-short strategy using news tones from BERT, FinBERT, RoBERTa, Baichuan, ChatGLM, and the ensemble model are 47.45%, 51.08%, 54.88%, 66.54%, 37.37%, and 69.21%, respectively, all with significant t-statistics. If we use return forecasts, the returns range between 35.51% and 45.64%. The abnormal returns obtained after risk adjustments are still around 41.50% to 71.72% (38.18% to 49.10%) when using news tones (return forecasts). That is to say, the signals (news tone or return forecasts) generated by LLMs using Chinese texts significantly predict future stock returns in China. The LLMs predictive power is more pronounced for stocks with stronger arbitrage limits and less transparent information environments, e.g., stocks with lower market capitalization, shorting volume, institutional and state ownership.

Regarding the second research question on the speed of news assimilation, and the speed of price reactions to the news, the efficient market hypothesis of Fama (1970) posits that expected returns are driven primarily by unpredictable news events, as this information is swiftly and fully reflected in asset prices. However, market inefficiencies, such as arbitrage constraints and opaque information environments, may prevent instantaneous news assimilation. Our empirical results show that the news tone can predict returns over the next 2 days, indicating that the information contained in news tone is absorbed into prices in 2 days. We also look into whether price moves before news is announced, which would indicate either a leakage or an expected move. Interestingly, prices start to move one day before the news announcement, indicating a price leakage considering that most of the news are un-prescheduled.

There might be two channels to explain the LLM signal's return predictive power. First, the LLMs might help to discover important information about the firm, which is not yet fully incorporated into prices, and this information is the driver for LLM signals' predictive power. We refer to this possibility as the "information" hypothesis. Second, it is possible that the LLMs generate signals for trading, and the trading creates short-term price pressure and demand for liquidity, which leads to short-term prediction. We refer to this possibility as the "trading" hypothesis. Empirically, we find that the LLM signals contain information about firms' fundamental information, and they also predict short-term trading directions. That is, the LLM signals' predictive power for future return is consistent with both channels, where the channel related to firms' fundamental information is likely linked to longer term predictions, and the channel related to short-term liquidity generates short-term price movements.

Our study relates to the following strands of research. The first is textual analysis. A rich series of studies, including Tetlock (2014), Tetlock et al. (2008), Loughran and McDonald (2011), Jegadeesh and Wu (2013), Loughran and McDonald (2016), Gentzkow et al. (2019), and Ke et al. (2019), demonstrates the significant return forecasting power of textual sentiment using English text. Besides forecasting stock returns, text data can be used in predicting market volatility as in Manela and Moreira (2017), modelling business cycles in Bybee et al. (2023a), and proxying for latent ICPM state variables in Bybee et al. (2023b). We contribute to this strand of research by effectively extracting information from Chinese news articles.

The second strand of studies adopts AI in asset pricing, with the most advanced technique now being the LLM. Gu et al. (2020) is among the pioneering works that employ machine learning methods to outperform traditional linear models. As AI technology evolves, studies rapidly emerge to examine the economic implications of LLMs for financial markets, mostly in the context of US stock market. Chen, Kelly, and Xiu (2023) investigate how various LLMs can be used to process news articles and forecast stock returns in US and other 15 foreign markets. Specifically looking into ChatGPT, Lopez-Lira and Tang (2023) interpret news headlines to predict cross-sectional stock returns, Chen, Tang, Zhou and Zhu (2023) process Wall Street Journal to predict aggregate market movements, and Beckmann et al. (2024) investigate unusual patterns in earnings calls and examine stock market reactions. Using the BERT model, Kim and Nikolaev (2023) extract operating profitability from annual reports and examine asset pricing factor models. Together, this emerging literature suggests LLMs can extract informative signals about asset prices and impact information dissemination and market efficiency. In China, however, there remains a blank of utilizing cutting-edge LLMs to extract news information and predict future stock returns. To date, sentiment analysis in China, including Li et al. (2019) and Jiang et al. (2021), relies largely on simplistic NLP methods that sacrifice contextual meaning. Our study is the first to adopt a representative series of Chinese LLMs to process the rich news information and incorporate China's unique market environment. We contribute to understanding cross-sectional stock returns and news information efficiency in China. We also demonstrate how modern NLP complements traditional methods in the field of finance.

The remainder of this paper is organized as follows. Section 2 introduces the data and methodology. Section 3 presents the main empirical results of predicting future returns using LLMs. Section 4 provides further discussion of the economic mechanism. Finally, Section 5 concludes the paper. Detailed discussions of methodology and additional results are provided in the Appendixes.

2. Data and Methodology

2.1 Data

2.1.1 Chinese News Articles

The Chinese news text dataset is obtained from ChinaScope SmarTag, a leading database provided by Mikuang Technology. We obtain extensive coverage of over 28 million news articles from January 2008 to December 2023, spanning traditional financial media, government sources, and WeChat official accounts across 8,732 sites.

Following the procedure in Chen, Kelly, and Xiu (2023), we apply the following filters to the news data. First, only articles that can be mapped to stocks are retained. Many macroeconomic, industry, and entertainment news items that are difficult to associate directly with stocks are removed. Second, to ensure each article is related to one stock, we remove articles tagged with more than one stock. When the article mentions multiple stocks, we cannot label its representation with a single stock-level return, which creates accuracy issues for subsequent econometric modeling in the training sample. Third, we only retain news related to the Chinese A-share stock market. News on stocks from the US, Hong Kong, or other countries is removed. Finally, we require that stocks have available returns data, so unlisted and delisted stocks are removed.

As shown in Panel A of Table 1, the initial dataset contained 28,259,596 raw articles. A total of 8,372,112 articles tagged with a single stock are retained after filtering out 19,887,484 non-stock-specific or multiple-stock-mentioned articles. A total of 2,233,748 articles referencing A-shares are retained after removing 6,138,364 news items on B/H-share and other international stocks. Finally, 2,193,371 articles merged with the return data are preserved after removing 40,377 articles without returns. These approximately 2.2 million articles serve as the main sample news for the following empirical analysis.

2.1.2 Other Data Items

We merge the news data with price, trading and accounting data from WIND and CSMAR. Following Ke et al. (2019) and Chen, Kelly, and Xiu (2023), we compute daily return using open-to-open returns, which is defined as $ret_{i,t} = \frac{OpenPrc_{i,t+1}}{OpenPrc_{i,t}} - 1$, where $OpenPrc_{i,t}$ means the adjusted price of stock *i* on day *t* at the market open.

Chinese stock market opens at 9:30 a.m., and closes at 3:00 p.m. Here we mainly focus on the open-to-open return for two reasons. First, to align with the Chinese "T+1" trading system (i.e., stocks purchased on one day cannot be sold until the next trading day). As shown in Appendix Figure A1, most news is released after the 3 p.m. market close in China. Thus, reacting to Day T news requires buying at the opening of Day T+1 and selling at the opening of Day T+2. Second, to allow for the constraint in trading timeliness. As discussed by Chen, Kelly, and Xiu (2023), overnight news can be challenging to act on before morning opening, as this is the earliest time most traders can access the market. Most funds (except high-frequency funds) are unlikely to continuously adjust positions intraday owing to investment styles and process constraints.

We also use trading size data to compute the order imbalance (*Oib*) of four groups of trading for each stock on each day. CSMAR integrates trade-by-trade data, and provides four groups of trades varying in trade size for each stock on each day: trades with small, medium, large, and extra-large sizes. Specifically, if the size of a trade is lower than 50,000 CNY, then it is a small-size trade. If the size of a trade is higher than or equal to 50,000 CNY but lower than 200,000 CNY, then it is a medium-size trade. If the size of a trade is higher than or equal to 200,000 CNY but lower than one million CNY, then it is a large trade. If the size of a trade is greater than or equal to 1 million CNY, then it is an extra-large trade. The order imbalance for a specific group of trades G for stock *i* on day *t* is calculated as the ratio of that group's buy minus sell to the sum of buy and sell volume. That is, $Oib_{i,t,G}=\frac{Buy_{i,t,G}-Sell_{i,t,G}}{Buy_{i,t,G}+Sell_{i,t,G}}$.

We also obtain accounting data items including stock's market capitalization (the product of the closing price and total A shares outstanding), earnings-to-price ratio (EP ratio, which is the ratio of the most recently reported quarterly net profit excluding non-recurrent gains/losses over last month-end's market capitalization.), turnover (daily share trading volume divided by tradable shares outstanding), and quarterly earnings data (the most recently reported net profit excluding nonrecurrent gains/losses). We then calculate quarterly unexpected earnings ($SUE_{i,t}$) for each firm *i* in each quarter *t*. Following Liu et al. (2019), *SUE* is calculated using a seasonal random walk, in which the year-over-year change in firm earnings is divided by the standard deviation of the previous eight quarters' year-over-year changes. That is, $SUE_{i,t} = \frac{\Delta_{i,t}}{\sigma(\Delta_i)}$ where $\Delta_{i,t}$ equals the year-over-year change in stock *i*'s quarterly earnings, and where $\sigma(\Delta_i)$ equals the standard deviation of $\Delta_{i,t}$ for the last eight quarters. Finally, we obtain stock ownership and state ownership data, where ownership refers to the proportion of shares held by a certain group of investors. The identification of state ownership follows Leippold et al. (2021).

2.2 Empirical Method

2.2.1 Econometric Modelling

Following Ke et al. (2019), our study has two key modelling objectives using financial news text: form news tones and construct return forecasts. Forming news tones allows us to assess the overall tones of news articles (i.e., whether they are positive or negative). Constructing return forecasts provides us with the forecasted returns associate with the news articles. Econometric modelling illuminates the statistical relationship between news and market movements. Our ultimate aim is to understand the economic implications of these text-based signals.

For forming news tones, we estimate a logistic model:

$$E\left(y_{i,t+1} \middle| x_{i,t}\right) = \sigma\left(x'_{i,t}\beta\right). \tag{1}$$

The labeled $y_{i,t}$ represents a dummy variable indicating whether the next day's return is positive or negative. If the stock return is positive (negative) the day after news publication, the article is labeled as having a positive (negative) tone, that is, *y=1* (*y=0*). Variable $x_{i,t}$ is the LLMs' article-level representation from news embedding, the derivation process of which is explained in following subsections. $\sigma(.)$ is the logistic function $\sigma(x) = \exp(x)/(1 + \exp(x))$. We estimate parameters β by minimizing the cross-entropy loss between the predicted and true class probabilities. With the sample estimate $\hat{\beta}$, we term $\sigma(x'_{i,t}\hat{\beta})$ the estimated news tone, ranging from [0,1], with values closer to one representing more positive tones. When news tone is higher than 0.5, the news is classified as positive, and when news tone is less than 0.5, the news is classified as negative.

For return forecast, we estimate a linear model:

$$E(r_{i,t+1}|x_{i,t}) = x'_{i,t}\theta \, , \tag{2}$$

where $r_{i,t+1}$ is the continuous next-day stock return from *t+1* open to *t+2* open, and $x_{i,t}$ is the LLMs' article-level representation. We estimate parameters θ by minimizing the mean squared error loss. With the sample estimate, $\hat{\theta}$, we define $x'_{i,t}\hat{\theta}$ as the return forecasts.

To address overfitting due to the high dimensionality of the text data, we regularize both models using ridge regression, as in Chen, Kelly, and Xiu (2023). We add an L2 penalty controlled by a tuning parameter α to shrink the model coefficients and prevent overfitting.[2] We determine

[2] An L2 penalty is a form of regularization that adds the sum of squared coefficients multiplied by a penalty parameter, α, to the objective function. α is the optimized hyperparameter that controls the strength of the penalty. By adding the penalty, the number of non-zero coefficients decreases and the phenomenon where a model learns the noise in the training data too well which results in poor generalization performance on unseen data is mitigated. In terms of optimizing α, 5-fold cross-validation means resampling where the original data is randomly partitioned into 5 equal subsamples, and the model is trained on 4 subsamples and evaluated on the remaining subsample, and this process is repeated 5 times with different test subsamples to obtain a cross-validated estimate.

the optimal α using 5-fold cross-validation with a grid search over the log range of 1e-5−1e5.

Our training sample spans 2008 to 2018, and testing sample spans 2019 to 2023. For model training, we utilize an expanding window approach, in which the in-sample training data grows annually. In the first iteration, the training window spans 2008-2018 (11 years) to get the first-round estimation of $\hat{\beta}$ and $\hat{\theta}$. The year 2019 is reserved for out-of-sample testing. In the next iteration, the training window expands to 2008−2019 (12 years), with the year 2020 set aside for testing. This continues for five iterations, expanding the training window by one additional year. With this expanding window approach, our out-of-sample test years range from 2019 to 2023. The expanding windows allow our models to accumulate more training data over time while still evaluating the performance on 5 years of out-of-sample data. This helps balance the model improvement from larger training sets with rigorous out-of-sample testing on data completely excluded from training.[3]

2.2.2 Large Language Models

We estimate six LLMs in total. The first model is BERT. The BERT model is a pretrained language model introduced by Devlin et al. (2018). It innovatively undertakes two unsupervised objectives: masked language modelling and next-sentence prediction. The masked language

The grid search is an exhaustive search over a geometric progression of α values spanning 5 orders of magnitude on the log scale to find the value that optimizes the cross-validation performance.

[3] Some may be concerned that our LLMs have learned from future news data during pre-training, and our current empirical results may not be out-of-sample predictions with look-ahead bias. We argue this is not a valid concern for two reasons. First, the basic model we use, Google's BERT, was fully trained and released publicly on November 3, 2018, using data only up to that date. Our out-of-sample prediction starts on January 1, 2019 after BERT's release; thus, look-ahead bias is not possible. Second, while some of our newer models (e.g., Baichuan and ChatGLM) may possibly have seen certain future news during pre-training, it only exposes models to the textual content of news articles, not the corresponding future stock returns or investor reactions. That is, the models are not informed by humans on whether each article was “good news” or “bad news” and do not learn to associate articles with returns and value judgements. Simply reading news content does not automatically imbue a model with forward looking bias.

modeling objective involves randomly replacing some input tokens with a special masking symbol. The model is then trained to predict the original vocabulary tokens for those masked positions, using the contextual information from the surrounding unmasked tokens. Meanwhile, the next sentence prediction objective trains the model to determine whether two given text segments maintain sequential coherence with respect to the original context from which they are extracted. BERT is pretrained on large volumes of text to teach the model relationships between words, thereby equipping it with knowledge that can be transferred to downstream tasks through fine-tuning. Its strength lies in the representativeness, as it acts as a benchmark for all subsequent LLMs. Its limitations include training with a relatively small batch size (further refined by RoBERTa) and few finance-domain contexts (improved by FinBERT). Specifically, we adopt the version of BERT optimized for the Chinese language.[4]

The second model is RoBERTa. The RoBERTa model is an optimized replication of pre-training BERT. It builds on key ideas from the original BERT but modifies hyperparameters, such as removing the next sentence prediction objective, training on longer sequences, and training on larger mini-batches and datasets. Together, Liu et al. (2019b) suggests these changes bring advantage of training stability and enhance the performance compared with BERT across NLP benchmarks. However, the weakness of lacking training on domain-specific data in financial context remains. We incorporate the RoBERTa model in our analysis to examine the economic gains from improving LLM's training stability. Specifically, we use the Chinese adaptation of the

[4] Specifically, we use the model bert-base-chinese from Hugging Face.

original RoBERTa, a large version of XLM-RoBERTa proposed by Conneau et al. (2020).[5]

The third model is FinBERT. The FinBERT model raised by Yang et al. (2020) is a financial domain-specific model based on BERT pre-trained parameters, fine-tuned on a large scale of financial corpora. We use an open-source Chinese pretrained FinBERT model, Valuesimplex FinBERT.[6] This is the first open-source Chinese model pre-trained on financial corpus containing 30 billion tokens from financial news, analyst reports, company announcements, and financial encyclopedia entries. Previous experiments show FinBERT has strengths in strong performance on downstream financial NLP tasks, such as financial sentiment analysis and relation extraction. Its potential weakness includes the number of parameters inherited from the original BERT, which may restrict its modeling capacity for highly complex and long contexts.

Since the release of ChatGPT in late 2022, there has been a rapid proliferation of research efforts and institutional investments aimed at developing sophisticated LLMs worldwide. We use following criteria and select another two models: (i) to be able to understand Chinese language, trained on Chinese corpora;[7] (ii) to obtain complete weight information, fully open-sourced; (iii) to ensure an accurate understanding of architecture, publicly available with detailed technical documentation; and (iv) high adoption among researchers and practitioners. ChatGLM and Baichuan are the last two LLMs we include which satisfy these criteria.

ChatGLM designed by Zeng et al. (2023) is an open bilingual language model with the Du et

[5] XLM-RoBERTa is a multilingual adaptation of RoBERTa pre-trained on 2.5 terabytes of filtered CommonCrawl data encompassing 100 languages, learning useful representations across multiple languages. Specifically, we use the model xlm-roberta-large from Hugging Face.

[6] Model sources from https://github.com/valuesimplex/FinBERT.

[7] Due to our criterion of selecting only Chinese-specific models trained on Chinese corpora, we did not include English-based models such as OPT (Zhang et al., 2022) and Llama (Touvron et al., 2022) in our study, despite their strong capabilities.

al.'s (2022) general language model (GLM) architecture. We adopt the most recent open-source version, the ChatGLM3-6B.[8] Its strength in generating human-like preferred responses comes from the massive 6.2 billion parameters, which are built on approximately one trillion tokens of bilingual Chinese-English training supplemented by supervised fine-tuning, self-supervised feedback, human-in-the-loop reinforcement learning, and other techniques. The potential weakness of ChatGLM, however, includes the general-purpose pretraining which may limit its capacity to capture financial predictive signals. Since March 2023, when the initial open-source version was publicly released, ChatGLM-6B has attracted over 10 million downloads on Hugging Face, demonstrating a high level of interest and usage.

The fifth model is Baichuan**.** The most up-to-date version, Baichuan-2 proposed by Yang et al. (2023), is a series of multilingual LLMs containing 13 billion parameters trained on 2.6 trillion tokens.[9] Unlike the GLM architecture used by ChatGLM, Baichuan-2 uses a model architecture consistent with LLAMA. That is, transformer's decoder-only architecture. Baichuan-2's advantages include matching or exceeding other open-source models of comparable sizes on public benchmarks, while the lack of training on finance-specific data still remains a drawback. The Baichuan series has attracted over 120 thousand downloads on Hugging Face since the initial open-source version's public release in June 2023, demonstrating its popularity.

The final model is the ensemble. To extract common signals, synthesize distinct predictors across models, and provide a unified economic interpretation, we employ an ensemble approach

[8] Specifically, we use the model THUDM/ChatGLM3-6b-base from Hugging Face.
[9] Specifically, we use the model baichuan-inc/Baichuan2-7B-Base from Hugging Face.

that aggregates information across LLMs. The ensemble method could be a powerful machine learning technique that combines the predictions of multiple models and improves the overall accuracy and robustness, as suggested by Dietterich (2000). Statistically, ensemble may improve the prediction accuracy by reducing the overall variance and canceling out individual biases. Economically, it may synthesize different aspects of future return-relevant text information that each model captures, reflecting different investor interpretations or firm fundamentals.

In our context, we construct an ensemble predictor by taking the average of the estimated news tones or return forecasts from the five individual LLMs. Simple averaging offers an easy, less capacity-consuming, and interpretable ensemble method. Though it does not necessarily outperform more complex approaches, its efficiency, simplicity, and ability to reduce variance as pointed out by Hansen and Salamon (1990) make it a valuable tool in the machine learning toolbox. Our purpose is not to design the best-performing model by inventing a complicated ensemble technique, but to show the potential value added by the aggregation process with a reasonable start.

The LLM takes several steps to derive the article-level representation, $x_{i,t}$, which is used as the key independent variable in Equation (1) and (2).[10] Given a tokenized article, a LLM maps each token to an embedding vector using the model's pre-trained embedding matrix. The total number of tokens that can be mapped, however, exists a limit for each model, and the dimensionality of vector varies. BERT-based models can process up to 512 tokens and encode them into 1,024-dimensional vectors. ChatGLM and Baichuan can handle around 8k and 4k tokens,

[10] We refer readers to Appendix 1 for details on tokenization, transformer architecture, pre-training, and fine-tuning.

with each token embedded in a 4,096-dimensional vector. Panel B of Table 1 reports the distribution of the number of tokens in Chinese news articles. Converted using model-specific tokenizers, the median of the number of tokens is around 220 for BERT-based models, 153 for Baichuan, and 161 for ChatGLM. Given the distribution of tokens, for articles with less than 512 tokens, we retain all tokens, whereas for articles exceeding 512 tokens, we use only the first 512 tokens. It maximizes the processing capacity of BERT-based models, and is consistent with Chen, Kelly, and Xiu (2023). Furthermore, around 93% (99%) of articles are under the threshold for BERT-based (Baichuan and ChatGLM) models. We then follow Chen, Kelly, and Xiu (2023) and take average of all token vectors to compute $x_{i,t}$. It can summarize full article content and represent the overall semantic information.

Besides LLMs, we also consider a traditional NLP method for comparison. Unlike LLMs which consider word order, syntax, and context, BOW initially proposed by Harris (1954) simply considers the occurrence and frequency of words to produce the tone for each news article. We follow Jiang et al. (2021) in constructing Chinese news tones using this BOW approach.[11] We attempt basic visualization to provide intuitions into what information this traditional method can capture. Appendix Figure A2 presents word clouds for the top and bottom scored news by the BOW. Panel A (B) shows commonly occurring terms in the highest (lowest) scored texts include "同比增长(亏损)", i.e., year-on-year growth (losses), among others. The word clouds provide directions of themes that may have been driving predictions in traditional textual analysis.

[11] The readers are referred to Appendix 2 for details of the calculations.

2.2.3 Performances of LLMs

We measure performances of LLMs using three approaches: out of sample fits, portfolio returns, and Fama-MacBeth regressions. For the out of sample fits, a true positive (true negative) prediction, *TP* (*TN*), occurs when a news tone > 0.5 (< 0.5) aligns with a realized positive (non-positive) return on the next day. False positives (FP) and false negatives (FN) represent complementary cases in which news tones do not match realized returns on the next day. We calculate the out-of-sample tone accuracy as the proportion of correct predictions in the testing sample: $Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$. We also calculate the out-of-sample cross-sectional correlations as the time-series average of the cross-sectional correlations between each model's return forecasts and the realized next-day returns, for each year in the testing sample.

For the portfolio approach, we adopt a trading timeline in which the opening at 9:30 a.m. to the next opening at 9:30 a.m. is considered a full trading day. Using all the news released during day *t*, we construct portfolios at the opening of day *t+1*. This trading timeline synchronizes with the "T+1" settlement cycle and considers the limit in intraday trading. When constructing portfolios, stocks are sorted by their news tones (or return forecasts) calculated from the LLMs. The top and bottom decile portfolios with the most and least positive news tones (or the highest and lowest return forecasts) are denoted as the long-leg (L) and short-leg (S), respectively. A zero-net investment L-S strategy refers to taking long (short) positions in the long-leg (short-leg). The portfolios are held for one day, and then rebalanced at the next opening on day *t+2* using all the news during day *t+1*. News released on non-trading days is postponed to the open of the next trading day, when it is first observable for trading decisions. With portfolios constructed, we examine their equal- and value-weighted annualized returns. Equal-weighting leans more towards small size stocks, which is economically meaningful in Chinese stock market with retail investors dominating the trading. If news tones and return forecasts can predict future returns, we expect a significantly positive annualized return for the L-S strategy.

To incorporate risk exposure, we also adjust the raw returns using the CH4 factor model of Liu et al. (2019a) which is tailored to the Chinese stock market. We estimate each portfolio's return series using the CH4 factor model and calculate the intercept term's coefficient (also referred to as "Alpha") and t-statistics. We use Newey and West (1987) to adjust the standard errors.

The third approach is the Fama-MacBeth (1973) regression. For each day *t*, we estimate the following cross-sectional specification:

$$ret_{i,t+1} = a_{0,t+1} + a_{1,t+1}X_{i,t} + a'_{2,t+1}Controls_{i,t} + u_{i,t+1}\ , \tag{3}$$

where $ret_{i,t+1}$ refers to the open-to-open returns of stock *i* at time *t+1*, and $X_{i,t}$ refers to news tones (*Tone*) or return forecast (*ER*) at time *t*. Control variables include previous open-to-open return (*LRet*) at time *t*, previous week open-to-open return (*Lwret*), previous month open-to-open return (*Lmret*), size (*Lsize*), EP ratio (*Lep*), and turnover (*Lturn*). We obtain the time series of the parameter estimates $\{a_{0,t+1}, a_{1,t+1}, a'_{2,t+1}\}$ from the cross-sectional regressions and conduct inferences on the mean and standard errors of these parameter estimates. Newey-West standard errors are adjusted using six lags. If news tones and predicted returns can forecast future stock price movements, we expect a significantly positive coefficient of a_1, the time-series average of $a_{1,t+1}$.

3. Model Fitting Performance and Return Prediction

In this section, we examine the model fitting performance and return predictability of the news tone and return forecast provided by the LLMs. We adopt three approaches: out-of-sample accuracy in Section 3.1, portfolio sorting in Section 3.2, and Fama-Macbeth regression in Section 3.3.

3.1 Out-of-Sample Accuracy

Table 2 Panel A presents the out-of-sample prediction accuracy by year for the LLMs. First, and most importantly, the six models consistently outperform random guessing (accuracy = 50%) over the sample period. Second, among all models under study, the BERT model delivers the highest overall accuracy of 52.71%, exceeding that of the RoBERTa model (52.63%), the ensemble model (52.58%), the FinBERT model (52.55%), the Baichuan model (52.32%), and the ChatGLM model (51.89%). Third, the accuracy of the BERT model peaks in the year 2022 at 54.16%, which is higher than all other models' accuracy in all out-of-sample years.

Table 2 Panel B reports the time-series average of the cross-sectional correlations between each model's return forecast and the realized next-day open-to-open return for each year in the testing sample. The correlations are positive for all models in all testing years and generally higher than 1%. Baichuan and the ensemble model have the highest overall correlation of 1.99% and 1.96%, followed by the FinBERT model (1.94%). RoBERTa, BERT, and ChatGLM's overall correlations are 1.80%, 1.62% and 1.52%, respectively.

3.2 Portfolio Returns

We first examine the return of portfolios sorted by news tones. The methodology is described in Section 2.2.3. In addition to LLMs, since the BOW model can also provide news tones, we include it to provide a direct comparison of return predictability between traditional NLP and state-of-the-art LLM methods.

In Figure 1, starting with the ensemble model, we compare the out-of-sample cumulative log returns of the long-short (L-S), long (L), and short (S) portfolios sorted by news tones. The solid and dashed lines represent equal-weighted (EW) and value-weighted (VW) portfolios, respectively. As the black solid line shows, the long-short strategy rapidly cumulates its return almost non-decreasingly throughout the testing sample, increasing from the log return of 0 to above 4, and outperforms the market represented by the yellow line. Therefore, the ensemble model is useful for extracting the tone of news articles in China, which takes time to be incorporated into prices.

In Figure 2, we report each LLM's cumulative log returns for EW L-S portfolios sorted by news tones. We also provide a comparison with the traditional BOW model. The results clearly indicate that (i) the ensemble model outperforms all other models, (ii) all LLMs outperform the traditional BOW model, and (iii) all models outperform the market. For ranking among LLMs, with the ensemble being the top performing model, the Baichuan model is the second-best, and the FinBERT and RoBERTa models steadily outperform the BERT and ChatGLM models throughout the testing sample. Overall, it is evident that adopting LLMs to capture the textual implications of Chinese language has great value to add to the investment industry.

Table 3 reports detailed returns on portfolios sorted by news tones. From Panel A where we examine raw returns, four key patterns emerge. First, the word-counting BOW model underperforms all LLMs by delivering the lowest annualized EW (VW) L-S return of 57.49% (24.96%) with a *t*-Stat of 7.73 (2.30), showing the superiority of LLMs in extracting useful information from text data beyond traditional NLP method. Second, across all LLMs and both weighting schemes, long legs consistently outperform short legs with the L-S strategies generating significantly positive annualized returns ranging from 64.28% (37.37%) to 89.88% (69.21%) when EW (VW), demonstrating the economic value of the tone signals provided by LLMs. Third, among all LLMs, the ensemble model delivers the highest returns. Specifically, taking the EW scheme for instance, ranking from the best performing LLM to the least, the L-S strategy has an annualized return of 89.88% with a *t*-Stat of 10.13 for the ensemble model, 84.40% with a *t*-Stat of 9.27 for the Baichuan model, 77.55% with a *t*-Stat of 8.99 for the FinBERT model, 74.26% with a *t*-Stat of 9.52 for the RoBERTa model, 65.26% with a *t*-Stat of 9.13 for the BERT model, and 64.28% with a *t*-Stat of 7.98 for the ChatGLM model. Similarly, for the VW results, ranking from the best performing LLM to the least, the L-S strategy has an annualized return of 69.21% with a *t*-Stat of 5.56 for the ensemble model, 66.54% with a *t*-Stat of 5.57 for the Baichuan model, 54.88% with a *t*-Stat of 4.93 for the RoBERTa model, 51.08% with a *t*-Stat of 4.52 for the FinBERT model, 47.45% with a *t*-Stat of 4.60 for the BERT model, and 37.37% with a *t*-Stat of 3.42 for the ChatGLM model. Fourth, the EW L-S returns exceed their VW counterparts, suggesting broader predictability benefits in small-cap stocks. Taking the ensemble model for instance, its EW L-S return of 89.88% is higher than its VW counterparts of 69.21%.[12]

In Table 3 Panel B, we conduct risk-adjustment to the raw returns in Panel A using the CH4

[12] The Sharpe ratio (SR) result of each model's L-S strategy is provided in Appendix Table A1. Ranking the LLMs from delivering the highest SR to the lowest, the EW L-S strategy has a SR of 4.93 for the ensemble model, 4.73 for the RoBERTa model, 4.62 for the Baichuan model, 4.46 for the FinBERT model, 4.40 for the BERT model, and 3.97 for the ChatGLM model. All LLMs' SRs are higher than that of the BOW model (3.88).

factor model of Liu et al. (2019). Four key patterns are identified. First, all the models deliver significantly positive alphas. Second, the BOW model again underperforms all other LLMs by yielding the lowest alpha of 59.22% (25.97%) when EW (VW). Third, the ensemble model has the highest alpha and outperforms all other LLMS. Specifically, ranking from the top performing LLM to the bottom, the ensemble model has a significant alpha of 92.57% with a *t*-stat of 11.92, Baichuan's alpha being 87.09% with a *t*-stat of 10.91, FinBERT's alpha being 78.95% with a *t*-stat of 9.68, RoBERTa's alpha being 75.22% with a *t*-stat of 10.25, BERT's alpha being 66.13% with a *t*-stat of 9.83, and ChatGLM's alpha being 67.23% with a *t*-stat of 9.41. Finally, the alphas become lower in magnitude yet still statistically significant when switching from the EW to the VW scheme.[13]

While the news tone shows both economically and statistically significant return predictability, return forecasts from news articles may also contain valuable information beyond merely predicting the direction of price movement. Table 4 presents the portfolio performances sorted by the cross-section of the LLMs' return forecasts using the methodology described in Section 2.2.3. Panel A shows three key patterns of portfolios' raw returns. First, similar to Table 3, across all the LLMs and weighting schemes, the long legs consistently outperform the short legs. For instance, ranking from the best performing model to the least, the EW L-S strategies generate significantly positive annualized returns of 79.53% for the ensemble model (*t*-Stat of 9.23), 76.17%

[13] In Appendix Table A2, we examine the heterogeneity of news types. We mainly examine three types of news: firm announcements (48.27% of all news articles), operation news (21.74%) and equity news (17.72%), which together constitute 87.73% of all news articles and are representative of the news diversity. The news types are categorized by the SmarTag database. We find that EW L-S portfolios sorted on firm announcement news tones have the highest annualized (risk-adjusted) return of 101.04% (103.13%), exceeding that of the equity news, 89.24% (91.93%), and that of the operation news, 49.43% (50.88%).

for Baichuan (*t*-Stat of 9.36), 73.21% for FinBERT (*t*-Stat of 9.40), 66.96% for ChatGLM (*t*-Stat of 7.93), 66.35% for RoBERTa (*t*-Stat of 8.51), and 56.49% for the BERT (*t*-Stat of 7.46). Second, the EW method outperforms the VW, signaling higher return predictability within small-cap stocks, where analyzing public information is less transparent and straightforward for the dominant retail investors. Third, unlike in Table 3, where ChatGLM delivers the lowest annualized L-S raw returns among all LLMs, its performance improves when more text features are absorbed. Specifically, ChatGLM increases its annualized return from 64.28% (*t*-Stat=7.98) in Table 3 to 66.96% (*t*-Stat=7.93), which is higher than RoBERTa's 66.35% (*t*-Stat=8.51) and BERT's 56.49% (*t*-Stat=7.46). Finally, the results are similar and robust when returns are risk-adjusted using the CH4 model in Panel B. Overall, return predictability using text features extends beyond news tones. [14]

In summary, we provide evidence that (i) forming L-S equity portfolios based on news tones and return forecasts provided by LLMs can yield significant economic gains in the Chinese stock market, and (ii) LLMs outperform traditional NLP method in return prediction. These results leave implication for the efficient market hypothesis that when we generally consider the Chinese A-share market to be increasingly efficient over time, news sentiment information takes at least one day to be incorporated into prices. The speed of such incorporation is discussed in Section 4.2.

3.3 Fama-MacBeth Regressions

As the third approach, we estimate Fama-MacBeth (1973) regressions to examine the return predictive power of news tones and return forecasts instead of long-short portfolio sorts. The

[14] In Appendix Table A3, we provide the Sharpe ratios of EW and VW L-S portfolios and their long and short legs.

methodology is described in Section 2.2.3. Table 5 presents the results of estimating equation (3).

Table 5 demonstrates consistent return predictability across LLMs. In Panel A, the positive and significant coefficients of the news tones indicate that higher tones predict greater future returns. Specifically, the coefficient for *Tone* is 0.0179 (1.79% daily) with a *t*-Stat of 11.12 for the ensemble model, 0.0178 (1.78% daily) with a *t*-Stat of 11.10 for BERT, 0.0166 (1.66%) with a *t*-Stat of 11.00 for RoBERTa, 0.0155 (1.55%) with a *t*-Stat of 9.86 for FinBERT, 0.0112 (1.12%) with a *t*-Stat of 9.72 for Baichuan, and 0.0068 (0.68%) with a *t*-stat of 9.50 for ChatGLM. Panel B shows similar results for the LLMs' return forecasts. Both panels suggest the significantly return predictability of text-based signals, even after controlling for characteristics that might influence future returns. Overall, the LLMs effectively forecast cross-sectional returns based on news text.

4. News Assimilation and Market Efficiency

In this section, we further discuss the economic implications of our main results. In Section 4.1, we investigate the return predictability for firms with different costs of arbitrage and information efficiency. In Section 4.2, we examine the speed of news assimilation and price responses. In Section 4.3, we empirically test the information and trading channels to better explain the return predictability.

4.1 Firms with Different Costs of Arbitrage and Information Efficiency

In this section, we aim to understand the heterogeneity of stock return predictability using LLMs, for stocks with different costs of arbitrage and information environments. We consider four economically important characteristics as proxies. (1) Market capitalization. Diamond and Verrecchia (1991) and Bhushan (1989) suggest that firm size may proxy for information environment, where smaller firms tend to be opaquer with higher information asymmetry. (2) Short volume. Short-selling constraints are more stringent in China than in other countries in terms of both eligibility restriction and low lendable supply. Only a portion of stocks are approved for margin trading and shorting. For instance, in 2023, only 2,770 of over 4,000 listed stocks could be shorted. Even among the eligible stocks, the supply of stocks available for shorting is quite low. In our sample, more than 60% of stocks have zero monthly shorting volumes. Studies including Saffi and Sigurdsson (2011) suggest that zero shorting activity may proxy for recent high shorting costs and limits to arbitrage which allows overpricing from sentiment momentum to persist longer without being arbitraged away. (3) Institutional ownership. Retail investors dominate the investor structure and contribute 80% of total trading in the Chinese stock market, according to Jones et al. (2023). Lower institutional ownership may proxy for a less professional investor base and more noise trading. This may prevent efficient assimilation of news information and incur sentiment distortions. (4) State ownership. State-owned enterprises in China may have low operational efficiency due to bearing nonprofitable social responsibilities. However, Jiang and Kim (2020) suggest that they have better corporate governance than non-state-owned enterprises, which helps reduce information asymmetry and improve information transparency. News reports have a more significant supplementary effect on information disclosure of non-state-owned stocks with lower information quality.

Table 6 presents the performance of portfolios sorted by either news tones or return estimates within each characteristic subgroup. "Large-Cap" ("Small-Cap") denotes the higher-than-median (lower-than-median) size group. Stocks with non-zero (zero) shorting volume over the past month are denoted as “Nonzero Shorting” (“Zero Shorting”). “High Institution” (“Low Institution”) denotes the higher-than-median (lower-than-median) institutional ownership group. “State Owned” (or “Non State Owned”) is an indicator for whether the stock is state-owned or not.

We start from Panel A of Table 6, where news tones are used to sort portfolios. Regarding size, the EW L-S strategy conducted on small-cap stocks has a higher annualized return of 109.12% with a *t*-Stat of 8.78, compared to 73.14% (*t*-Stat of 7.71) for large-cap stocks. These results confirm the previous findings that EW outperform VW strategies. Regarding short selling, stocks with zero shorting volume in the past month have higher EW L-S returns (106.22% annualized) than stocks with positive shorting volume during the past month (69.58% annualized). Regarding the role of institutional investors, stocks with lower institutional ownership have higher EW L-S returns of 97.12% annualized than those with lower retail ownership (78.96% annualized). Finally, non-state-owned stocks have an annualized EW L-S strategy return of 93.60%, which is higher than the 53.93% return for state-owned stocks. Panel B presents the robust results when sorted by return forecasts.

In summary, we show that stocks with lower caps, short-selling activities, institutional and state ownership enable higher return predictability of LLMs’ signals. Therefore, LLMs are helpful in capturing under-processed information in public news, especially for firms with higher costs of arbitrage and less efficient information environment, contributing to overall market efficiency.

4.2 Speed of News Assimilation and Price Response

The efficient market hypothesis of Fama (1970) posits that expected returns are driven primarily by unpredictable news events, as this information is swiftly and fully reflected in asset prices. However, market inefficiencies, such as arbitrage constraints and opaque information environments, may prevent instantaneous news assimilation. In this section, we test the speed of news assimilation and price response to examine the information efficiency of public news.

We first examine the speed of news assimilation in terms of when initiating a strategy utilizing an announced news could earn positive returns. Figure 3 compares the average one-day holding period returns to the news trading strategy based on the ensemble model as a function of when the trade is initiated. We consider daily open-to-open returns initiated one to four days after announcements. Figure 3 plots the average annualized EW CH4 risk-adjusted returns on the long-short (L-S), long-leg (L), and short-leg portfolios (S), with 99% confidence intervals given by the shaded regions in the top panel and the corresponding SRs in the bottom panel. The top panel shows the following major findings. The L-S strategy possesses significant and considerable positive returns on Day 0, Day 1, and Day 2, but starts to diminish to zero on Day 3. Meanwhile, the return on Day 1 is less than half of that on Day 0, stressing the need for a timely strategy to interpret news tones. In terms of the risk-adjusted returns in the bottom panel, the SR of the L-S strategy is the highest on Day 0 and Day 1, followed by Day 2, but quickly decreases from the beginning of Day 3, which provides additional insights into the speed of news assimilation. Generally, from the right half of Figure 3, we find that news is primarily assimilated into prices before the third trading day following its announcement.

Another interesting pattern emerging from Figure 3 is that the news assimilation process seems to start one day before the announcement. From the left-half figures where we plot the four-day window before announcement, Day -1 has significantly positive L-S returns and SR, which are even higher than those on Day 2. Our results seem to suggest that relevant information is priced before it becomes public.

Motivated by the above findings, we investigate the speed of price response, particularly regarding whether the information content in a news article is already incorporated into prices before a news announcement. Figure 4 compares the out-of-sample EW cumulative log returns of the L-S portfolios sorted on news tones provided by the ensemble model, with each line representing the relative timing of portfolio construction. Note that constructing a portfolio based on news tone before the announcement would not constitute a tradable strategy because it is forward looking, but it would signal the association between the price movement one day before the news and news tone itself. We represent this association in the *Day -1* line. Similarly, while constructing the portfolio on a news day requires immediate transactions and is not always practical, such a strategy signals the association between news and returns on the same day. We denote this association by the *Day 0* line. An L-S portfolio constructed one day after announcement is denoted by the *Day +1* line.

We notice two patterns in Figure 4. First, trading on the news tone one day before the announcement cumulates substantial returns during the testing years. The cumulative log return increases from zero to nearly three, and are infrequently reversed throughout the years. The positive return indicates a leakage or an expected move. Considering that most news are announced without prescheduling, the information leakage hypothesis more likely holds. Second, trading on news tone on the same day as the news leads to greater profits than one day later (nearly threefold of log cumulative returns), meaning that prices already reflect a major bulk of information on news announcement days, and less so one day later.[15]

In summary, for information efficiency, we find that news assimilation takes two days, and in the one-day horizon ahead of a public announcement, prices already respond.

4.3 Hypotheses

Two channels might help to explain the LLM signal's return predictive power. First, news may convey important fundamental information about firms' future performance, as suggested by Tetlock et al. (2008). The LLMs might help to discover this information, which is not yet fully incorporated into prices. This information is the driver for LLM signals' predictive power. We refer to this possibility as the "information" hypothesis. If this hypothesis holds, we expect that the news tones extracted by LLMs can positively predict firms' future earnings surprises.

H1: Information Hypothesis. News tones convey firms' fundamental information which is not yet reflected in stock prices but captured by LLMs.

Second, it is possible that the LLMs generate signals for trading, and the trading creates short-term price pressure. Specifically, when a piece of news conveying an optimistic (pessimistic) tone becomes public, investors who wish to immediately profit from the news before price fully adjusts may swiftly submit aggressive buying (selling) orders. This behavior could have a temporary upward (downward) price impact according to Griffiths et al. (2000), leading to the short-term

[15] Figure 3 and Figure 4 suggest an early price response in advance of news announcement. If such early response is correct (i.e., consistent with the direction of the announced news), then a strategy trading on the news tones would become less profitable as information already prices in to certain extent. However, if such early response is incorrect (i.e., contrary to the direction of the announced news), then such strategy should become more profitable and economically crucial in terms of arbitraging mispricing. In Appendix Table A4, we find the long strategy performs better when stocks are trending down compared with trending up, and the short strategy performs better when stocks are trending up compared with trending down. Thus, when price early responds inconsistently with the news direction, there is more room for arbitrage and profitability of LLM strategies.

return predictability. Such trading can also create demand for liquidity as aggressive trades rush in to trade in the same direction. We refer to this possibility as the “trading” hypothesis. If this hypothesis is true, then for aggressive trades that have a higher demand for timely trading using the LLMs’ signals, future trading directions should be positively related to previous news tones extracted by the LLMs. However, for relatively less aggressive trades, future trading directions could be negatively related to previous news tones if they act as liquidity providers.

H2: Trading Hypothesis. LLMs generate signals for trading, where the more aggressive trades tend to follow while the less aggressive trades act as counterparties. The aggressive trades bring temporary price pressure, leading to the short-term return predictability.

We report the results for *H1* in Table 7. Following Tetlock et al. (2008), we adopt OLS regression to predict future earnings surprises using previous news tones. The dependent variable is future quarterly unexpected earnings ($SUE_{i,t}$) for each firm *i* in each quarter *t*, defined in Section 2.1.2. The key independent variable is the news tone provided by the ensemble model that occurs one trading day before the announcement of firm earnings. We include lagged control variables similar to those in Table 5. Following Froot (1989), standard errors are clustered by calendar quarter.

From Table 7, we notice that the news tone extracted by LLMs has a significantly positive coefficient of 5.08 with a *t*-stat of 9.10. That is, the higher the news tones, the higher the future earnings surprise. The results suggest that news tone captures firms’ fundamental information, which is then gradually incorporated into prices, leading to the return predictability of news tones. Thus, the information hypothesis, *H1*, is supported.

Next, we examine *H2* in Table 8. We adopt the Fama-Macbeth (1973) method to predict next-day trading directions using previous news tones, similar to equation (3). The dependent variables are the next-day order imbalances for the four types of trade: $Oib_{i,t,Small}$, $Oib_{i,t,Medium}$, $Oib_{i,t,Large}$, and $Oib_{i,t,ExtraLarge}$. Variable definitions are provided in Section 2.1.2. The key independent variable is the news tone provided by the ensemble model. In addition to the control variables mentioned in Section 2.2.3, we additionally control for the previous day's order imbalance (*Loib*) to allow for trading persistence, following Jones et al. (2023).

We use trade size as a proxy for trading aggressiveness. We assume that the larger the trade size, the more aggressive the trade. For the most aggressive trades with extra-large sizes of more than one million CNY, news tone has a significantly positive coefficient of 7.20 with a *t*-Stat of 7.94. Meanwhile, for similarly aggressive large trades ranging between 200,000 and 1 million CNY, news tone has a significantly positive coefficient of 4.98 with a *t*-Stat of 5.35. However, for the least aggressive trades with trade sizes lower than 50,000 CNY, news tone has a significantly negative coefficient of -1.89 with a *t*-Stat of -3.84. That is, news tones appear to positively drive the future direction of the aggressive trades: the more positive (negative) the news tones are, the more aggressive trades net buy (sell) and demand liquidity. The least aggressive type of trade provides liquidity and acts as a counterparty. Overall, these results support the trading hypothesis of *H2*.[16]

Overall, we find supportive evidence for both information and trading hypotheses, where the

[16] In Appendix Table A5, we show that the return predictability of news tones does not reverse in the longer horizon of 8 weeks.

news tones extracted by LLMs contain information about firms' fundamental and predict short-term trading directions.

5 Conclusion

This study provides novel evidence of LLMs' effectiveness for stock return prediction in the Chinese stock market using news text. First, we establish the predictive power of news tones from LLMs, with the ensemble model generating the best market-beating cumulative returns. The significant L-S portfolio alphas demonstrate market inefficiency for arbitrage opportunities. Moreover, all LLMs outperform traditional NLP method such as BOW. Robustly, the return forecasts provided by LLMs can also significantly predict future returns. Second, in terms of information efficiency, return predictability is more pronounced for stocks with higher costs of arbitrage and less efficient information environment, news is assimilated into prices within two days, and the price drift begins one day before the public release. Finally, both the information and trading hypotheses are supported to help explain return predictability.

Overall, our results highlight the need to adopt advanced LLMs for text analysis in the Chinese stock market. The expressive capabilities of LLMs arise from pretraining on massive text corpora, which enables them to learn language patterns transferable to downstream tasks. Our findings reveal the value of this transferability when fine-tuned for finance in China's unique linguistic and market environment. As future research continues to enrich LLMs with domain-specific data, their financial applications are expected to grow.

This study serves as a springboard for an exciting research agenda at the intersection of AI and finance in China. With advanced LLMs tailored to Chinese text, this largely understudied area is primed for rapid development. Our analysis may motivate creative applications of LLMs and spur advances in data-rich and computationally intensive text methods for investment research. More broadly, our study exemplifies the potential of artificial intelligence to extract insights from complex text data and enhance decision-making for investors in the Chinese stock market.

References

Beckmann, Lars, Heiner Beckmeyer, Ilias Filippou, Stefan Menze, and Guofu Zhou, 2024, Unusual Financial Communication - Evidence from ChatGPT, Earnings Calls, and the Stock Market, *SSRN Electronic Journal*.

Bhushan, Ravi, 1989, Firm Characteristics and Analyst Following, *Journal of Accounting and Economics* 11, 255–274.

Bybee, Leland, Bryan T. Kelly, Asaf Manela, and Dacheng Xiu, 2023, Business News and Business Cycles, *The Journal of Finance*, *forthcoming*.

Bybee, Leland, Bryan T. Kelly, and Yinan Su, 2023, Narrative Asset Pricing: Interpretable Systematic Risk Factors from News Text, *Review of Financial Studies, forthcoming*.

Carpenter, Jennifer N., Fangzhou Lu, and Robert F. Whitelaw, 2021, The Real Value of China's Stock Market, *Journal of Financial Economics* 139, 679–696.

Chen, Yifei, Bryan Kelly, and Dacheng Xiu, 2023, Expected Returns and Large Language Models, *SSRN Electronic Journal*.

Chen, Jian, Guohao Tang, Guofu Zhou, and Wu Zhu, 2023, ChatGPT, Stock Market Predictability and Links to the Macroeconomy, *SSRN Electronic Journal.*

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov, 2020, Unsupervised Cross-lingual Representation Learning at Scale, *arXiv preprint arXiv:1911.02116*.

Cui, Yiming, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang, 2019, Pre-Training with Whole Word Masking for Chinese BERT, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29, 3504–3514.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 2018, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *arXiv preprint arXiv:1810.04805*.

Diamond, Douglas W., and Robert E. Verrecchia, 1991, Disclosure, Liquidity, and the Cost of Capital, *The Journal of Finance* 46, 1325–1359.

Dietterich, Thomas G., 2000, Ensemble Methods in Machine Learning, Multiple Classifier Systems. Lecture Notes in Computer Science (Springer Berlin Heidelberg, Berlin, Heidelberg).

Du, Zhengxiao, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang, 2022, GLM: General Language Model Pretraining with Autoregressive Blank Infilling, *arXiv preprint arXiv:2103.10360.*

Fama, Eugene F, 1970, Efficient capital markets: A review of theory and empirical work, *The Journal of Finance* 25, 383–417.

Fama, Eugene F., and James D. MacBeth, 1973, Risk, Return, and Equilibrium: Empirical Tests, *Journal of Political Economy* 81, 607-636.

Froot, Kenneth A., 1989, Consistent Covariance Matrix Estimation with Cross-Sectional Dependence and Heteroskedasticity in Financial Data, *The Journal of Financial and Quantitative Analysis* 24, 333.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy, 2019, Text as Data, *Journal of Economic Literature* 57, 535–574.

Griffiths, Mark D, Brian F. Smith, D. Alasdair S. Turnbull, and Robert W. White, 2000, The costs and determinants of order aggressiveness, *Journal of Financial Economics* 56, 65-88.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu, 2020, Empirical Asset Pricing via Machine Learning, *The Review of Financial Studies* 33, 2223–2273.

Hansen, Lars Kai, and Peter Salamon, 1990, Neural Network Ensembles, *IEEE transactions on pattern analysis and machine intelligence* 12, 993–1001.

Harris, Zellig S., 1954, Distributional Structure, *Word* 10, 146–162.

Jegadeesh, Narasimhan, and Di Wu, 2013, Word Power: A New Approach for Content Analysis, *Journal of Financial Economics* 110, 712–729.

Jiang, Fuwei, Lingchao Meng and Guohao Tang, 2021, Media Textual Sentiment and Chinese Stock Return Predictability, *China Economic Quarterly* 21, 1323-1344. *(in Chinese)*

Jiang, Fuxiu, and Kenneth A Kim, 2020, Corporate Governance in China: A Survey, *Review of Finance* 24, 733–772.

Jones, Charles M., Shi, Donghui, Zhang, Xiaoyan and Zhang, Xinran, 2023, Retail Trading and Return Predictability in China, *Journal of Financial and Quantitative Analysis, forthcoming.*

Ke, Zheng Tracy, Bryan Kelly, and Dacheng Xiu, 2019, Predicting Returns with Text Data, *SSRN Electronic Journal*.

Kim, Alex G., and Valeri V. Nikolaev, 2023, Profitability Context and the Cross-Section of Stock Returns, *SSRN Electronic Journal*.

Kudo, Taku, and John Richardson, 2018, SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing, *arXiv preprint arXiv:1808.06226*.

Leippold, Markus, Qian Wang, and Wenyu Zhou, 2022, Machine learning in the Chinese stock market, *Journal of Financial Economics* 145, 64–82.

Li, Jia, Yun Chen, Yan Shen, Zhuo Huang, and Jingyi Wang, 2019, Measuring China’s Stock Market Sentiment, SSRN Electronic Journal.

Liu, Jianan, and Robert F. Stambaugh, and Yu Yuan, 2019, Size and value in China, *Journal of Financial Economics* 134, 48-69.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, 2019, RoBERTa: A Robustly Optimized BERT Pretraining Approach, *arXiv preprint arXiv:1907.11692*.

Lopez-Lira, Alejandro, and Yuehua Tang, 2023, Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models, *SSRN Electronic Journal*.

Loughran, Tim, and Bill Mcdonald, 2011, When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks, *The Journal of Finance* 66, 35–65.

Loughran, Tim, and Bill Mcdonald, 2016, Textual Analysis in Accounting and Finance: A Survey, *Journal of Accounting Research* 54, 1187–1230.

Manela, Asaf, and Alan Moreira, 2017, News implied volatility and disaster concerns, *Journal of Financial Economics* 123, 137–162.

Newey, Whitney K., and Kenneth D. West, 1987, A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix, *Econometrica* 55, 703-708.

Saffi, Pedro AC, and Kari Sigurdsson, 2011, Price efficiency and short selling, *The Review of Financial Studies* 24, 821–852.

Song, Changcheng, 2019, Financial Illiteracy and Pension Contributions: A Field Experiment on Compound Interest in China, *The Review of Financial Studies* 33, 916–949.

Tetlock, Paul C., 2007, Giving Content to Investor Sentiment: The Role of Media in the Stock Market, *The Journal of Finance* 62, 1139–1168.

Tetlock, Paul C., 2014, Information Transmission in Finance, *Annual Review of Financial Economics* 6, 365–384.

Tetlock, Paul C., Maytal Saar-Tsechansky, and Sofus Macskassy, 2008, More Than Words: Quantifying Language to Measure Firms' Fundamentals, *The Journal of Finance* 63, 1437–1467.

Titman, Sheridan, Chishen Wei, and Bin Zhao, 2022, Corporate Actions and the Manipulation of Retail Investors in China: An Analysis of Stock Splits, *Journal of Financial Economics* 145, 762–787.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample, 2023, LLaMA: Open and Efficient Foundation Language Models, *arXiv preprint arXiv:2302.13971*.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin, 2017, Attention is all you need, *Advances in neural information processing systems* 30.

Yang, Aiyuan, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, JunTao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu, 2023, Baichuan 2: Open Large-scale Language Models, *arXiv preprint arXiv:2309.10305*.

Yang, Yi, Mark Christopher Siy UY, and Allen Huang, 2020, FinBERT: A Pretrained Language Model for Financial Communications, *arXiv preprint arXiv:2006.08097*.

Zeng, Aohan, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Peng Zhang, Yuxiao Dong, and Jie Tang, 2023, GLM-130B: An Open Bilingual Pre-trained Model, *arXiv preprint arXiv:2210.02414*.

Zhang, Susan, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam

Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer, 2022, OPT: Open Pre-trained Transformer Language Models, *arXiv preprint arXiv:2205.01068*.

Table 1. Summary Statistics of Chinese News Articles

In this table, we report the summary statistics of Chinese news articles. Our sample period is from January 2008 to December 2023. Our sample stocks are A-share stocks listed on Shanghai Stock Exchange and Shenzhen Stock Exchange. News Articles are in Chinese and are obtained from ChinaScope SmarTag database. Panel A presents remaining sample size after each filter applied on the news articles. Column "Number of Articles Retained" presents remaining sample size while column "Number of Articles Filtered Out" presents sample size filtered out. Row "Raw Articles" presents the numbers of available articles from the ChinaScope SmarTag database. Row "Articles Tagged with Single Stock Code" presents the number of articles tagged with a single stock. Row "Articles Tagged with A-share Stock Event" presents the number of articles tagged with A-share stock events. Row "Articles with Returns" presents the number of remaining articles after matching returns data. In Panel B, we further report the summary statistics of the number of characters and tokens in the filtered sample. Row "# of Characters" report the percentiles of the number of characters in the raw article. Rows "# of BERT Tokens", "# of FinBERT Tokens", "# of RoBERTa Tokens", "# of Baichuan Tokens", and "# of ChatGLM Tokens" report the percentiles of the number of tokens converted from news text using model specific tokenizer.

Panel A. Sample Size and Filtering

	Number of Articles Retained	Number of Articles Filtered Out
Raw Articles	28,259,596	/
Articles Tagged with Single Stock Code	8,372,112	19,887,484
Articles Tagged with A-share Stock Event	2,233,748	6,138,364
Articles with Returns	2,193,371	40,377

Panel B. Characters/Tokens in Chinese News Articles

	1%	25%	50%	75%	99%
# of Characters	65	153	241	361	877
# of BERT Tokens	61	141	223	334	816
# of FinBERT Tokens	62	143	224	335	817
# of RoBERTa Tokens	62	143	224	335	817
# of Baichuan Tokens	41	97	153	227	530
# of ChatGLM Tokens	45	104	161	237	545

Table 2. Out-of-Sample Prediction Accuracy and Cross-Sectional Correlations

In this table, we report the out-of-sample statistical performances of the BERT, FinBERT, RoBERTa, Baichuan, ChatGLM, and their ensemble model. In Panel A, under the sentiment classification task, we compute the classification accuracy of the news tone provided by each model for each year in the testing sample. In Panel B, under the return regression task, we calculate the time-series average of the cross-sectional rank correlations between the return forecasts and the future one day's open-to-open returns for each model and each year in the testing sample.

Panel A. Out-of-Sample Prediction Accuracy Under the Sentiment Classification Task

Year	BERT	FinBERT	RoBERTa	Baichuan	ChatGLM	Ensemble
2019	52.21%	52.21%	52.21%	52.49%	52.28%	52.29%
2020	51.45%	51.35%	51.45%	51.32%	50.48%	51.36%
2021	52.67%	52.56%	52.60%	52.39%	52.04%	52.62%
2022	54.16%	53.73%	53.87%	52.97%	52.84%	53.76%
2023	53.08%	52.90%	53.01%	52.40%	51.83%	52.90%
Overall	52.71%	52.55%	52.63%	52.32%	51.89%	52.58%

Panel B. Out-of-Sample Cross-Sectional Correlations Under the Return Regression Task

Year	BERT	FinBERT	RoBERTa	Baichuan	ChatGLM	Ensemble
2019	1.65%	1.79%	1.69%	2.08%	2.03%	2.01%
2020	1.68%	2.24%	2.07%	2.27%	1.59%	2.23%
2021	2.30%	2.80%	2.89%	2.56%	2.02%	2.72%
2022	1.51%	1.59%	1.31%	1.70%	1.60%	1.75%
2023	0.94%	1.26%	1.03%	1.31%	0.33%	1.07%
Overall	1.62%	1.94%	1.80%	1.99%	1.52%	1.96%

Table 3. Performance of Daily News Tone Portfolios in China

The table reports the performance of equal-weighted (EW) and value-weighted (VW) long-short (L-S) portfolios sorted on news tones and their long (L) and short (S) legs. The decile portfolios are built based on the traditional BOW model or LLMs, including BERT, FinBERT, RoBERTa, Baichuan, ChatGLM, and LLMs' ensemble model. In Panel A, "Ret" and "*t*-Stat" stand for each portfolio's annualized return and *t*-Statistics. In Panel B, "Alpha" and "*t*-Stat" stand for each portfolio's annualized CH4-adjusted return and *t*-Statistics.

Panel A. Raw Returns (Annualized) for Long and Short Portfolios

Model		EW		VW	
		Ret	*t*-Stat	Ret	*t*-Stat
BOW	L	27.09%	2.70	15.37%	1.46
	S	-30.40%	-2.51	-9.58%	-0.81
	L-S	57.49%	7.73	24.96%	2.30
BERT	L	26.51%	2.63	15.90%	1.59
	S	-38.75%	-3.21	-31.55%	-2.73
	L-S	65.26%	9.13	47.45%	4.60
FinBERT	L	34.28%	3.23	14.07%	1.28
	S	-43.28%	-3.30	-37.02%	-3.00
	L-S	77.55%	8.99	51.08%	4.52
RoBERTa	L	30.39%	2.94	21.30%	2.02
	S	-43.87%	-3.53	-33.58%	-2.83
	L-S	74.26%	9.52	54.88%	4.93
Baichuan	L	41.11%	3.80	27.74%	2.39
	S	-43.28%	-3.47	-38.79%	-3.35
	L-S	84.40%	9.27	66.54%	5.57
ChatGLM	L	31.21%	2.99	18.58%	1.67
	S	-33.07%	-2.84	-18.79%	-1.77
	L-S	64.28%	7.98	37.37%	3.42
Ensemble	L	40.38%	3.66	23.72%	2.13
	S	-49.50%	-3.88	-45.49%	-3.66
	L-S	89.88%	10.13	69.21%	5.56

Panel B. CH4-adjusted Returns (Annualized) for Long and Short Portfolios

Model		EW		VW	
		Alpha	*t*-Stat	Alpha	*t*-Stat
BOW	L	17.24%	4.58	8.76%	1.50
	S	-41.97%	-7.18	-17.20%	-2.17
	L-S	59.22%	8.61	25.97%	2.47
BERT	L	16.33%	3.74	9.18%	1.54

	S	-49.79%	-9.41	-38.22%	-5.28
	L-S	66.13%	9.83	47.40%	4.97
FinBERT	L	24.68%	5.09	7.67%	1.20
	S	-54.26%	-8.47	-43.77%	-5.61
	L-S	78.95%	9.68	51.44%	4.92
RoBERTa	L	20.32%	4.61	14.56%	2.23
	S	-54.89%	-9.42	-39.89%	-5.33
	L-S	75.22%	10.25	54.45%	5.22
Baichuan	L	32.42%	6.75	22.92%	3.66
	S	-54.68%	-8.64	-46.99%	-5.92
	L-S	87.09%	10.91	69.90%	6.52
ChatGLM	L	22.53%	4.96	13.60%	2.23
	S	-44.70%	-8.31	-27.90%	-4.13
	L-S	67.23%	9.41	41.50%	4.23
Ensemble	L	31.64%	6.44	18.12%	2.91
	S	-60.92%	-9.89	-53.60%	-6.50
	L-S	92.57%	11.92	71.72%	6.45

Table 4. Performance of Portfolios Sorted by Return Forecasts

The table reports the performance of equal-weighted (EW) and value-weighted (VW) long-short (L-S) portfolios and their long (L) and short (S) legs. The portfolios are built based on BERT, FinBERT, RoBERTa, Baichuan, ChatGLM, and the ensemble model, respectively, using LLMs' return forecasts as sorting variables. In Panel A, "Ret" and "*t*-Stat" stand for each portfolio's annualized return and *t*-Statistics. In Panel B, "Alpha" and "*t*-Stat" stand for each portfolio's annualized CH4-adjusted return and *t*-Statistics.

Panel A. Raw Returns (Annualized) for Long and Short Portfolios

Model		EW		VW	
		Ret	*t*-Stat	Ret	*t*-Stat
BERT	L	27.11%	2.45	19.39%	1.69
	S	-29.38%	-2.47	-16.13%	-1.48
	L-S	56.49%	7.46	35.52%	3.50
FinBERT	L	35.05%	3.24	20.00%	1.73
	S	-38.16%	-3.22	-25.64%	-2.29
	L-S	73.21%	9.40	45.64%	4.45
RoBERTa	L	28.14%	2.49	20.47%	1.76
	S	-38.21%	-3.23	-17.54%	-1.54
	L-S	66.35%	8.51	38.01%	3.68
Baichuan	L	37.49%	3.42	14.91%	1.33
	S	-38.68%	-3.13	-26.21%	-2.39
	L-S	76.17%	9.36	41.12%	4.06
ChatGLM	L	33.70%	3.05	15.75%	1.32
	S	-33.26%	-2.77	-19.76%	-1.81
	L-S	66.96%	7.93	35.51%	3.46
Ensemble	L	38.65%	3.46	14.66%	1.24
	S	-40.88%	-3.36	-26.85%	-2.52
	L-S	79.53%	9.23	41.51%	3.89

Panel B. CH4-adjusted Returns (Annualized) for Long and Short Portfolios

Model		EW		VW	
		Alpha	*t*-Stat	Alpha	*t*-Stat
BERT	L	18.08%	3.56	13.79%	2.03
	S	-41.18%	-8.13	-24.39%	-3.65
	L-S	59.25%	8.19	38.18%	3.85
FinBERT	L	26.31%	5.20	14.97%	2.23
	S	-50.36%	-9.99	-34.12%	-5.02
	L-S	76.67%	10.71	49.10%	4.94
RoBERTa	L	18.98%	3.54	14.70%	2.27

	S	-50.13%	-10.25	-25.34%	-3.70
	L-S	69.11%	9.36	40.04%	4.00
Baichuan	L	29.16%	5.31	10.59%	1.69
	S	-50.74%	-9.55	-34.98%	-5.44
	L-S	79.90%	10.80	45.57%	5.04
ChatGLM	L	25.08%	4.44	10.54%	1.53
	S	-44.93%	-8.48	-27.96%	-4.30
	L-S	70.01%	8.86	38.51%	3.97
Ensemble	L	30.33%	5.20	10.19%	1.49
	S	-53.08%	-10.06	-35.18%	-5.28
	L-S	83.41%	10.42	45.37%	4.64

Table 5. Return Prediction Using Fama-MacBeth Regressions

This table reports the robustness return prediction results using Fama-Macbeth regression. The dependent variable is the next day open-to-open return. In Panel A, the key independent variable, *Tone*, is the news tone extracted by LLM. In Panel B, the key independent variable, *ER*, is the return forecast estimated by LLM. LLMs include BERT, FinBERT, RoBERTa, Baichuan, ChatGLM, and the ensemble model. We also include control variables in the regression, including previous open-to-open return (*LRet*), previous week open-to-open return (*Lwret*), previous month open-to-open return (*Lmret*), size (*Lsize*), EP-ratio (*Lep*) and turnover (*Lturn*). Newey-West adjusted standard errors are calculated using six lags.

Panel A. News Tones Predicting Returns

	BERT		FinBERT		RoBERTa		Baichuan		ChatGLM		Ensemble	
	Coef	*t*-Stat	*Coef*	*t*-Stat	*Coef*	*t*-Stat	*Coef*	*t*-Stat	*Coef*	*t*-Stat	*Coef*	*t*-Stat
Tone	0.0178	11.10	0.0155	9.86	0.0166	11.00	0.0112	9.72	0.0068	9.50	0.0179	11.12
Lret	-0.0147	-4.14	-0.0145	-4.04	-0.0147	-4.09	-0.0154	-4.29	-0.0154	-4.32	-0.0150	-4.19
Lwret	-0.0051	-3.05	-0.0050	-2.98	-0.0051	-3.04	-0.0051	-3.01	-0.0052	-3.09	-0.0051	-3.03
Lmret	-0.0005	-0.61	-0.0005	-0.63	-0.0005	-0.56	-0.0006	-0.70	-0.0007	-0.74	-0.0006	-0.73
Lsize	0.0000	-0.79	0.0000	-0.81	0.0000	-0.96	0.0000	-1.01	0.0000	-0.73	0.0000	-1.15
Lep	-0.0012	-0.33	-0.0021	-0.59	-0.0023	-0.63	-0.0026	-0.71	-0.0010	-0.26	-0.0036	-0.98
Lturn	-0.0124	-3.25	-0.0123	-3.25	-0.0127	-3.30	-0.0128	-3.41	-0.0125	-3.35	-0.0124	-3.30
Intercept	0.0019	5.08	0.0016	4.35	0.0017	4.66	0.0011	3.19	0.0008	2.31	0.0018	4.73
Adj.R2	0.04%		0.09%		0.07%		0.12%		0.09%		0.11%	

Panel B. Returns Forecasts Predicting Returns

	BERT		FinBERT		RoBERTa		Baichuan		ChatGLM		Ensemble	
	Coef	*t*-Stat	*Coef*	*t*-Stat	*Coef*	*t*-Stat	*Coef*	*t*-Stat	*Coef*	*t*-Stat	*Coef*	*t*-Stat
ER	0.4660	10.20	0.4844	10.73	0.4798	11.69	0.3218	11.22	0.1798	10.27	0.4643	11.82
Lret	-0.0150	-4.24	-0.0149	-4.24	-0.0149	-4.23	-0.0153	-4.34	-0.0153	-4.33	-0.0152	-4.33
Lwret	-0.0053	-3.17	-0.0054	-3.23	-0.0054	-3.22	-0.0054	-3.21	-0.0054	-3.23	-0.0055	-3.27
Lmret	-0.0005	-0.61	-0.0006	-0.70	-0.0006	-0.64	-0.0007	-0.76	-0.0006	-0.71	-0.0007	-0.74
Lsize	0.0000	-0.58	0.0000	-0.76	0.0000	-0.69	0.0000	-0.94	0.0000	-0.80	0.0000	-0.97

Lep	-0.0001	-0.02	-0.0009	-0.25	-0.0006	-0.18	-0.0016	-0.46	-0.0002	-0.06	-0.0016	-0.45
Lturn	-0.0129	-3.35	-0.0123	-3.19	-0.0128	-3.37	-0.0125	-3.29	-0.0121	-3.17	-0.0123	-3.23
Intercept	0.0003	0.78	0.0002	0.68	0.0003	0.89	0.0004	1.10	0.0004	1.18	0.0003	0.95
Adj.R2	0.06%		0.07%		0.07%		0.08%		0.08%		0.08%	

Table 6. Return Prediction Heterogeneity

This table reports the heterogeneity results of decile portfolios sorted by daily news tones or return forecasts for stocks with different costs of arbitrage and information efficiency. We first sort stocks into two subgroups based on each of the following four characteristics. (i) Market capitalization. "Large-Cap" (or "Small-Cap") denotes the higher-than-median (or lower-than-median) size subgroup. (ii) Shorting activity. "Nonzero Shorting" (or "Zero Shorting") is an indicator for whether the stock has nonzero (or zero) short-selling volume during the previous month. (iii) Institutional ownership. "High Institution" (or "Low institution") denotes the subgroup of stocks with higher-than-median (or lower-than-median) percentage of shares held by institutional investors. (iv) State-ownership. "State Owned" (or "Non State Owned") is an indicator for whether the stock is state-owned or not. Within each subgroup, we then sort on daily news tones (or return forecasts) based on the ensemble model of various LLMs in Panel A (or Panel B), and form equal-weighted decile portfolios. In Panel A, decile portfolios with the lowest (or highest) news tones are denoted by Short (S) (or Long (L)). In Panel B, decile portfolios with the lowest (or highest) return forecasts are denoted by Short (S) (or Long (L)). L-S denotes the long-short portfolios. Columns "Ret" and "*t*-Stat" stand for each portfolio's annualized return and *t*-Statistics.

Panel A. Heterogeneity When Sorted on News Tones

	Large-cap		Small-cap	
	Ret	*t*-Stat	Ret	*t*-Stat
L	32.12%	2.98	51.64%	3.85
S	-41.02%	-3.34	-57.48%	-3.72
L-S	73.14%	7.71	109.12%	8.78
	Nonzero Shorting		Zero Shorting	
	Ret	*t*-Stat	Ret	*t*-Stat
L	34.29%	2.98	48.28%	3.92
S	-35.29%	-2.87	-57.94%	-3.98
L-S	69.58%	6.90	106.22%	10.26
	High Institution		Low Institution	
	Ret	*t*-Stat	Ret	*t*-Stat
L	43.00%	3.76	39.29%	3.32
S	-35.96%	-2.80	-57.83%	-3.96
L-S	78.96%	7.29	97.12%	8.72
	State Owned		Non State Owned	
	Ret	*t*-Stat	Ret	*t*-Stat
L	18.70%	1.72	44.79%	3.87
S	-35.23%	-2.46	-48.81%	-3.64
L-S	53.93%	4.76	93.60%	9.15

Panel B. Heterogeneity When Sorted on Returns Forecasts

	Large-cap		Small-cap	

	Ret	*t*-Stat	Ret	*t*-Stat
L	21.89%	1.97	61.90%	4.33
S	-27.30%	-2.43	-51.83%	-3.51
L-S	49.19%	5.72	113.74%	8.90
	Nonzero Shorting		Zero Shorting	
	Ret	*t*-Stat	Ret	*t*-Stat
L	24.15%	2.10	52.24%	4.12
S	-22.87%	-1.88	-56.64%	-4.26
L-S	47.02%	4.46	108.88%	10.47
	High Institution		Low Institution	
	Ret	*t*-Stat	Ret	*t*-Stat
L	39.93%	3.54	36.90%	2.97
S	-29.09%	-2.43	-48.08%	-3.43
L-S	69.02%	7.21	84.98%	7.87
	State Owned		Non State Owned	
	Ret	*t*-Stat	Ret	*t*-Stat
L	23.24%	2.10	44.99%	3.62
S	-27.27%	-2.00	-45.67%	-3.76
L-S	50.50%	4.66	90.65%	8.61

Table 7. Earnings Surprise Prediction Using News Tones

This table reports the OLS regression results for predicting future earnings surprises using previous news tones. The dependent variable is the future quarterly unexpected earnings (SUE) for each firm in each quarter. Following Liu et al. (2019), SUE is calculated as the year over year change in earnings divided by the standard deviation of previous eight quarters' year over year changes. The key independent variable, *Tone*, is the news tone provided by the ensemble model of various LLMs. The timing of news tones is chosen to be the previous trading day of the firm's earnings announcement. We also include lagged control variables in the regression, including previous open-to-open return (*LRet*), previous week open-to-open return (*Lwret*), previous month open-to-open return (*Lmret*), size (*Lsize*), EP-ratio (*Lep*) and turnover (*Lturn*). Following Froot (1989), standard errors are clustered by calendar quarter.

Dep.Var	Next-day SUE	
	Coef	*t*-Stat
Tone	5.08	9.10
Lret	1.76	4.96
Lwret	1.35	6.80
Lmret	0.71	5.77
Lsize	0.02	14.09
Lep	0.86	1.86
Lturn	-0.79	-1.58
Intercept	0.53	8.56
Adj.R2	5.87%	

Table 8. Trading Order Imbalance Prediction Using News Tones

This table reports the Fama-Macbeth regression results for predicting next-day trading order imbalances using previous day news tones. The dependent variables are the next-day order imbalances for four types of trades: trades with small trade sizes, *Oib(Small)*, medium sizes, *Oib(Medium)*, large sizes, *Oib(Large)*, and extra-large sizes, *Oib(ExtraLarge)*. To be specific, if the size of a trade is lower than 50,000 CNY, then we identify such trade as a small-size trade. If the size of a trade is higher or equal to 50,000 CNY but lower than 200,000 CNY, then we identify such trade as a medium-size trade. If the size of a trade is higher or equal to 200,000 CNY but lower than one million CNY, then we identify such trade as a large-size trade. If the size of a trade is higher or equal to 1 million CNY, then we identify such trade as an extra-large-size trade. We assume that the higher the trade sizes, the more aggressive these trades are. Trade size data are obtained from CSMAR database. Order imbalance for a specific type of trade is calculated as the ratio of that type's buy minus sell over the sum of that type's buy and sell volume. The key independent variable, *Tone*, is the news tone provided by the ensemble model of various LLMs. We also include lagged control variables in the regression, including previous day order imbalance (*Loib*), previous open-to-open return (*LRet*), previous week open-to-open return (*Lwret*), previous month open-to-open return (*Lmret*), size (*Lsize*), EP-ratio (*Lep*) and turnover (*Lturn*). Newey-West adjusted standard errors are calculated using six lags.

Dep.Var	Next-day Oib(Small)		Next-day Oib(Medium)		Next-day Oib(Large)		Next-day Oib(ExtraLarge)	
	Coef	*t*-Stat	*Coef*	*t*-Stat	*Coef*	*t*-Stat	*Coef*	*t*-Stat
Tone	-1.89	-3.84	-0.26	-0.48	4.98	5.35	7.20	7.94
Loib	0.20	68.75	0.12	36.36	0.05	19.78	0.01	5.22
Lret	-5.90	-6.94	5.11	4.80	8.34	4.99	0.92	0.48
Lwret	1.71	3.98	6.47	11.83	7.88	9.85	-1.68	-2.25
Lmret	1.65	6.04	3.84	12.91	4.56	9.15	0.18	0.42
Lsize	0.06	7.89	0.17	16.76	0.16	13.82	-0.02	-1.84
Lep	12.14	6.66	10.94	5.51	-2.22	-0.87	-13.48	-4.42
Lturn	2.58	2.32	3.09	2.49	-6.24	-2.85	-5.95	-2.68
Intercept	-1.33	-11.84	-4.64	-30.84	-5.04	-19.45	0.99	6.30
Adj.R2	3.89%		1.56%		0.16%		0.00%	

Figure 1. One-day-ahead Portfolio Performance Based on Ensemble Model

This figure compares the out-of-sample cumulative log returns of portfolios sorted on news tones based on the ensemble model of various LLMs. The black, blue, and red colors represent the long-short (L-S), long (L), and short (S) portfolios, respectively. The solid and dashed lines represent equal-weighted (EW) and value-weighted (VW) portfolios, respectively. The yellow solid line represents the A-share market portfolio.

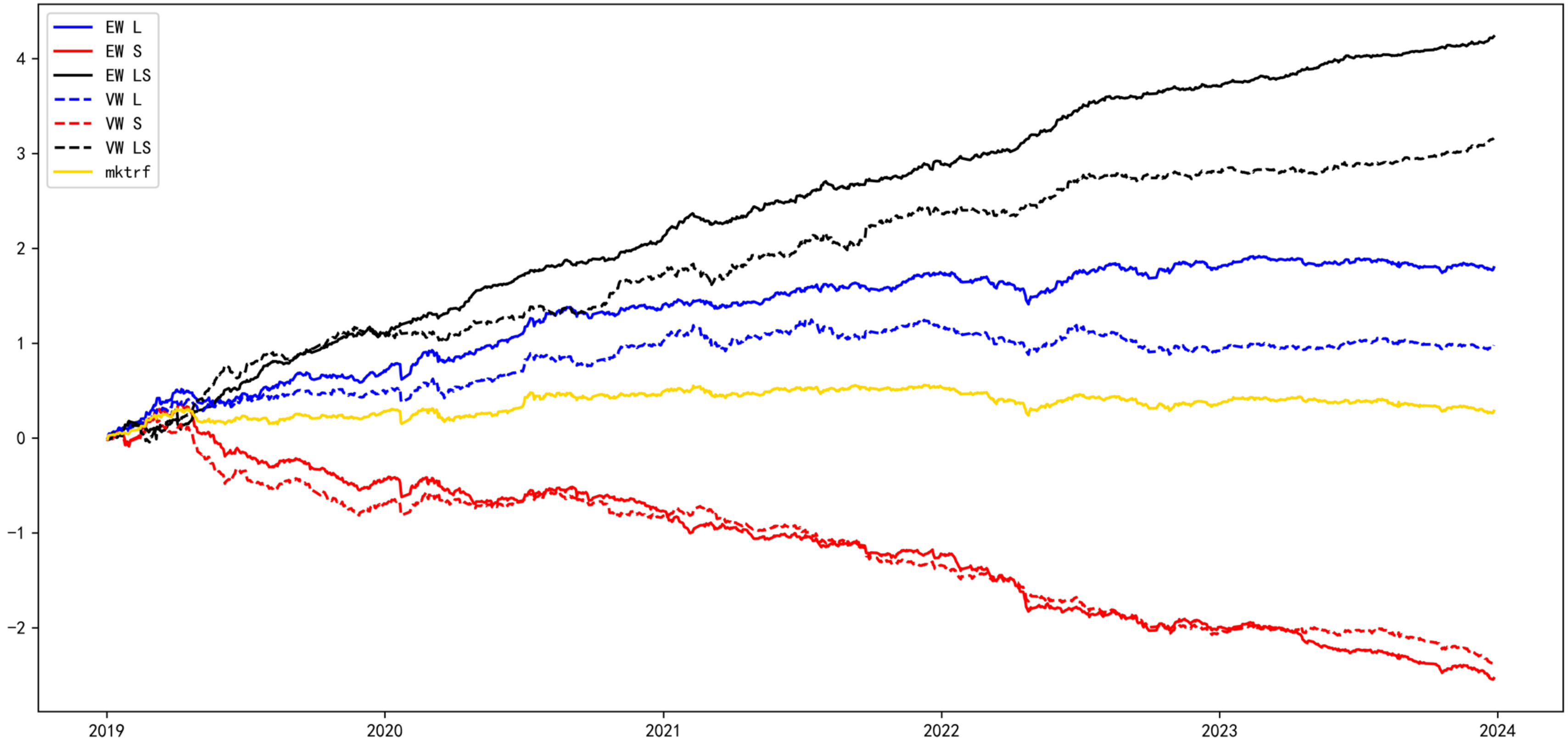


Figure 2. Performance of Equal-Weighted Portfolios Based on Various Models

This figure plots the cumulative log returns for equal-weighted long-short portfolios sorted on news tones. The portfolios are built based on BOW, BERT, FinBERT, RoBERTa, Baichuan, ChatGLM, and the ensemble model, respectively. "Mkt" represents the cumulative A-share market return.

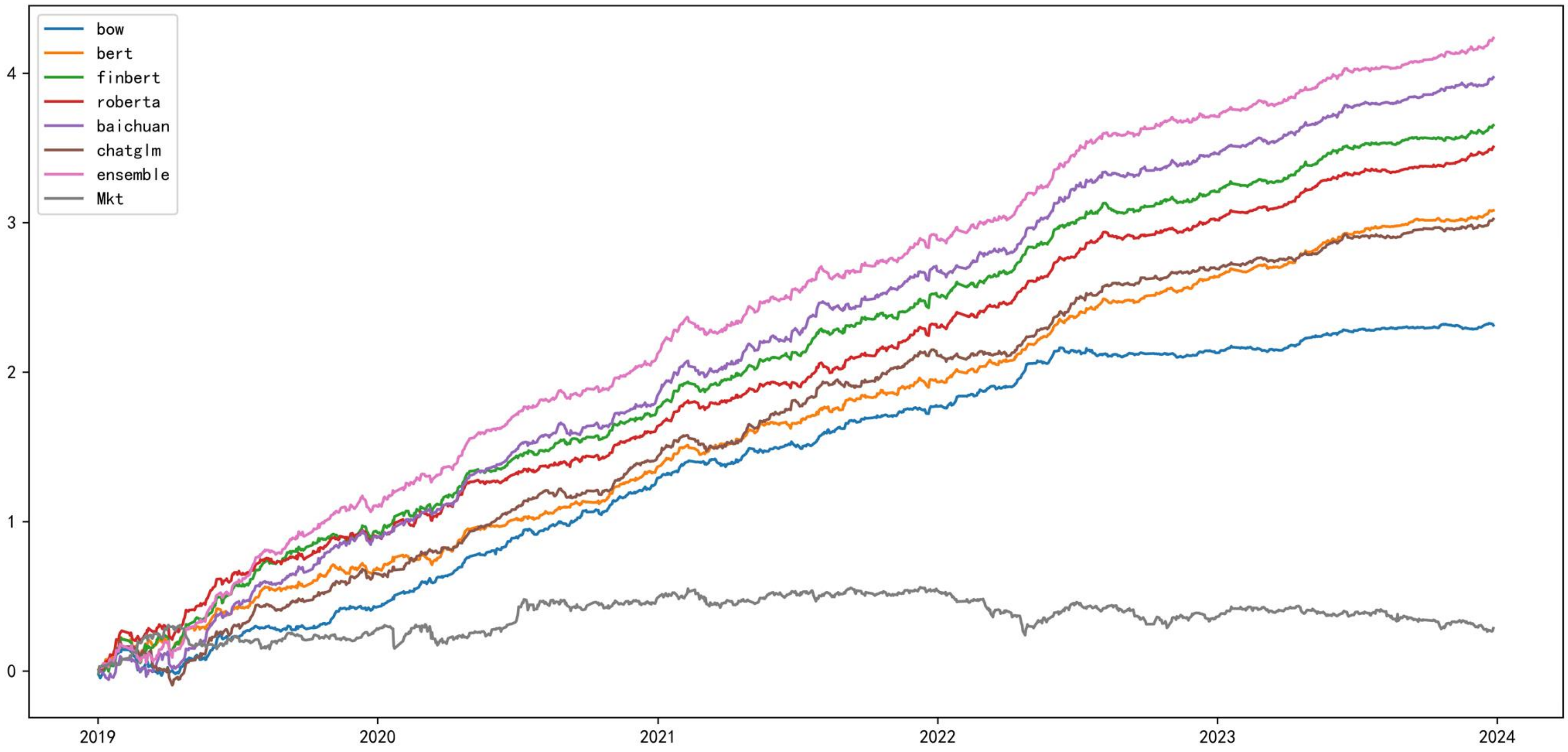


Figure 3. Speed of News Assimilation

This figure compares average one-day holding period returns to the news trading strategy based on the ensemble model of various LLMs, as a function of when the trade is initiated. We consider daily open-to-open returns initiated from one to four days following the announcement. We also examine the open-to-open returns in the four-day window ahead of the announcement. We report the average annualized equal-weighted CH4 risk adjusted returns on the long-short (L-S) portfolio, long-leg (L) portfolio and short-leg portfolio (S), with 99% confidence intervals given by the shaded regions on the top panel, as well as the corresponding Sharpe ratios on the bottom panel.

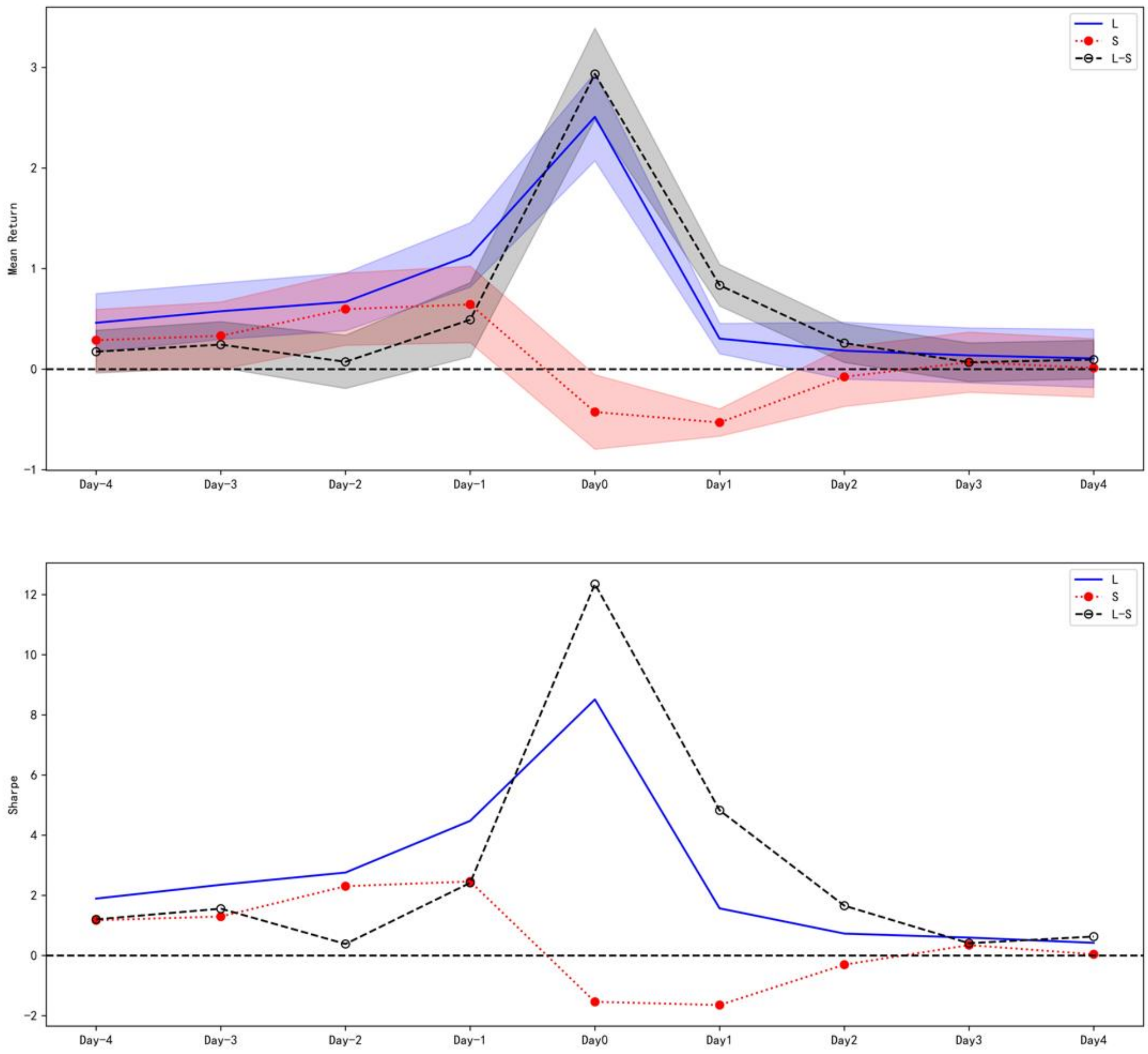
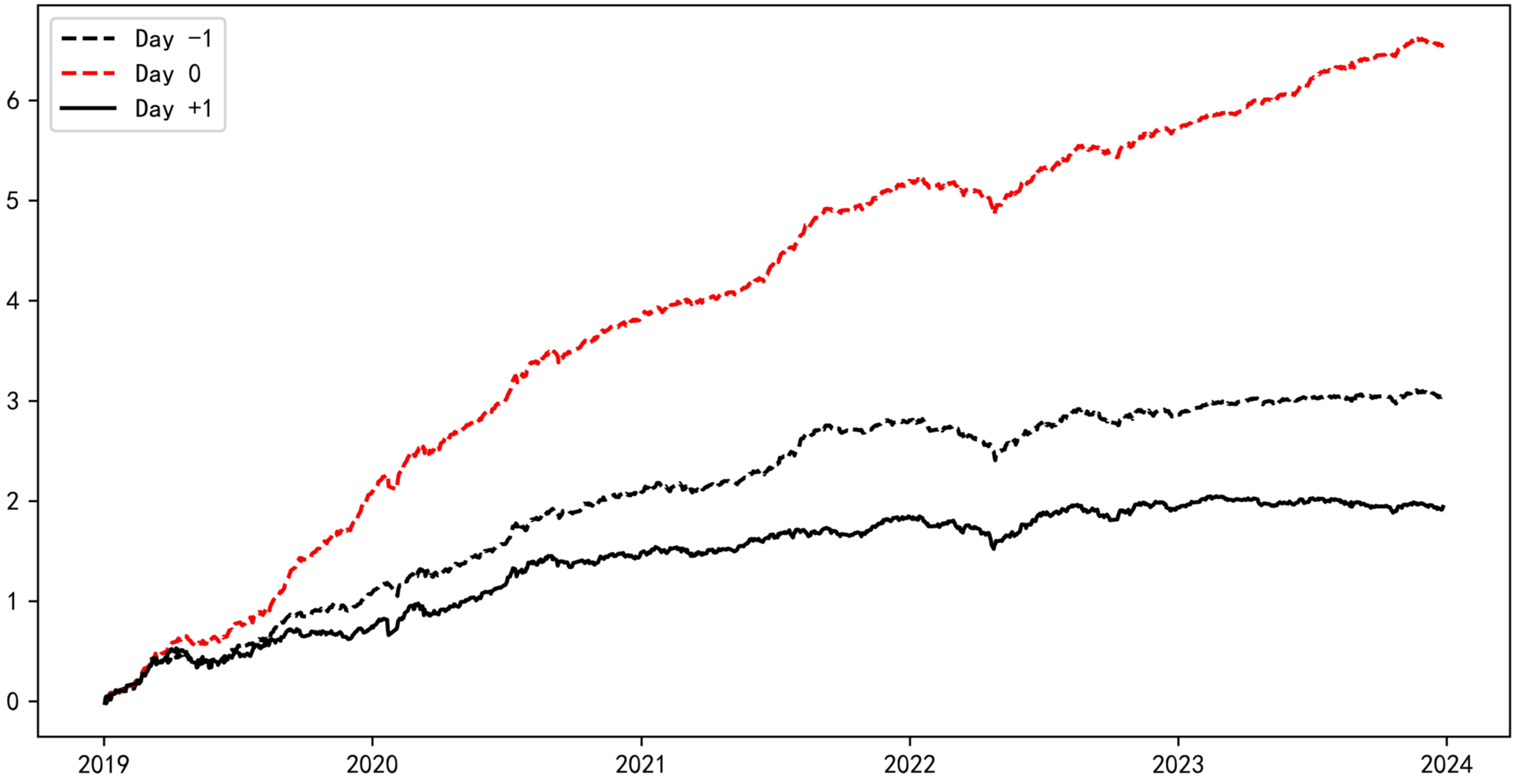


Figure 4. Price Response on Days −1, 0, and +1

This figure compares the out-of-sample equal-weighted cumulative log returns of long-short portfolios sorted on news tones provided by the ensemble model of various LLMs. The Day −1 line (dashed black line) shows the association between news and returns one day prior to the news; the Day 0 line (dashed red line) shows the association between news and returns on the same day; and the Day +1 line (solid black line) shows the association between news and returns one day later.



Appendix 1. Details on Large Language Models

Texts in Chinese, or any other language, consist of word sequences and semantics that are neither easily measurable nor comparable as numerical data do. Earlier studies including Tetlock (2007) and Loughran and McDonald (2011) extract signals from financial texts using simple statistical approaches, such as manually-defined financial dictionaries. Yet the high dimensionality and complex nature of language prevent traditional methods from fully capturing the richness of context. Recently, machine learning and deep learning techniques emerge as handful tools, among which LLMs offer the most advanced capabilities. LLMs adopt tokenization algorithms that effectively deal with word segmentation. They obtain rich language understanding through deep contextualized embeddings that retain semantics, word order, and cross-word relationships, and are pretrained on massive text corpora using deep neural networks. Fine-tuning further adapts the LLMs to specific downstream objectives suitable for financial analysis, such as the econometric modelling in Section 2.2.1. In this Appendix, we describe the above procedures in more details.

A1.1 Tokenization

In any NLP framework, contextualized representations originate from tokenization. The broken-down unit of text is referred to as a token, which can take the form of character, word, or sub-word, reflecting different tokenization algorithms.

A key challenge in Chinese tokenization, compared to that in English, is disambiguating words with ambiguous boundaries. While English words have explicit word boundaries with spaces, a sentence in Chinese does not separate words explicitly. Therefore, an accurate Chinese word segmentation must identify the word and phrase boundaries by incorporating the surrounding contextual semantics and syntax. LLMs employ the SentencePiece tokenization technique of Kudo and Richardson (2018) that can learn the optimal segmentation from training data. Superior to previous tokenization methods (e.g., dictionary-based), SentencePiece can automatically construct sub-word units from the text, effectively representing out-of-vocabulary words not seen during training. This improves the model's generalization capability for open vocabularies. Furthermore, the generated sub-words are smaller than words from traditional tokenization algorithms. This can better mitigate data sparsity, capture the compositional patterns between words, and improve the quality of extracted contextualized representations.

A1.2 Transformer Architecture

The transformer architecture is a neural network architecture first proposed by Vaswani et al. (2017), which is now commonly adopted in NLP. It employs an "encoder-decoder" structure and relies solely on attention mechanisms, discarding recurrence and convolutions entirely. This brings two key advantages over previous sequence transduction models: parallelization and long-range dependencies.

Specifically, the transformer encoder maps an input sequence to a continuous representation by applying multiple layers of multi-headed self-attention. Self-attention allows each position in the sequence to attend to all other positions and compute a representation that aggregates information from the entire sequence. Multi-headed attention splits this computation into multiple sub-spaces, providing multiple "representations of the sequence" which allows the model to jointly attend to information from different representation sub-spaces at different positions.

The transformer decoder, on the other hand, generates an output sequence by masking future positions, preventing leftward information flow, and preserving auto-regressive generation. It stacks multiple layers of multi-headed self-attention, followed by multi-headed attention over the encoder outputs, which enables each position in the decoder to make use of the full context from the complete input sequence.

A1.3 Pre-training

Pre-training is another common approach in NLP. The LLMs are first pre-trained on a large corpus of text in an unsupervised manner to learn useful linguistic representations before being fine-tuned on downstream tasks. Popular pre-training models, taking the BERT for instance, push the edge across many NLP benchmarks by pre-training deep bidirectional representations from the large corpora.

Pre-training provides two main advantages: (1) it allows models to learn universal language representations from massive amounts of unlabeled data, and (2) it enables transfer learning by initializing models with these pre-training parameters for improved performance on tasks with limited labeled data.

A1.4 Fine-Tuning

Fine-tuning refers to initializing a model with pre-training parameters and then training it on labeled data from downstream tasks.

Fine-tuning adapts the LLMs to fit new objectives by using minimal task-specific parameters. This enables models to build on existing knowledge from pre-training while customizing to new objectives with limited labeled data. Fine-tuning often achieves significant performance gains compared to training on downstream tasks from scratch.

Appendix 2. Details on Bag-of-Words (BOW) Model

In this appendix, we describe the process of deriving news tones using the BOW model following Jiang et al. (2021).

The first step involves word segmentation using the *jieba* toolkit for Chinese text processing. Next, we import the open-source Chinese financial sentiment dictionary by Jiang et al. (2021),[1] and perform a lookup in the dictionary for each token obtained from the word segmentation. The third step is to calculate the news tone for a news article. We adopt the TFIDF method following Loughran and McDonald (2011). The news tone equals the sum of weights of positive words minus the sum of weights of negative words. The weight calculation method is $w_{i,j} = \begin{cases} \frac{1+\log(tf_{i,j})}{1+\log(a)} log\frac{N}{df_i} & if\ tf_{i,j} \geq 1 \\ 0 & otherwise \end{cases}$, where $tf_{i,j}$ represents the number of times word i appears in news article j, a represents the average number of words per news article, N represents the total number of news articles, and df_i represents the number of news articles that contain word i.

By following this process, we are able to construct the BOW news tone for each news article, which serves as a benchmark for informativeness comparison with the news tone extracted by LLMs.

[1] https://github.com/MengLingchao/Chinese_financial_sentiment_dictionary

Appendix 3

Appendix Table A1. Sharpe Ratios of Daily News Tone Portfolios in China

The table reports the Sharpe Ratios of equal-weighted (EW) and value-weighted (VW) long-short (L-S) portfolios sorted on news tones and their long (L) and short (S) legs. The decile portfolios are built based on the traditional BOW model or LLMs, including BERT, FinBERT, RoBERTa, Baichuan, ChatGLM, and the LLMs' ensemble model. "Ret", "Std", and "SR" stand for each portfolio's annualized return, standard deviation, and Sharpe Ratio, respectively.

Model		EW			VW		
		Ret	Std	SR	Ret	Std	SR
BOW	L	27.09%	22.81%	1.19	15.37%	23.05%	0.67
	S	-30.40%	24.47%	-1.24	-9.58%	24.79%	-0.39
	L-S	57.49%	14.81%	3.88	24.96%	21.07%	1.18
BERT	L	26.51%	22.86%	1.16	15.90%	22.89%	0.69
	S	-38.75%	24.12%	-1.61	-31.55%	24.77%	-1.27
	L-S	65.26%	14.82%	4.40	47.45%	21.81%	2.18
FinBERT	L	34.28%	23.59%	1.45	14.07%	24.14%	0.58
	S	-43.28%	24.95%	-1.73	-37.02%	26.28%	-1.41
	L-S	77.55%	17.37%	4.46	51.08%	24.24%	2.11
RoBERTa	L	30.39%	23.21%	1.31	21.30%	23.33%	0.91
	S	-43.87%	24.39%	-1.80	-33.58%	26.33%	-1.28
	L-S	74.26%	15.71%	4.73	54.88%	23.54%	2.33
Baichuan	L	41.11%	23.99%	1.71	27.74%	26.11%	1.06
	S	-43.28%	24.37%	-1.78	-38.79%	25.16%	-1.54
	L-S	84.40%	18.25%	4.62	66.54%	26.12%	2.55
ChatGLM	L	31.21%	23.37%	1.34	18.58%	26.00%	0.71
	S	-33.07%	23.75%	-1.39	-18.79%	23.86%	-0.79
	L-S	64.28%	16.19%	3.97	37.37%	23.71%	1.58
Ensemble	L	40.38%	24.27%	1.66	23.72%	25.88%	0.92
	S	-49.50%	24.81%	-2.00	-45.49%	26.02%	-1.75
	L-S	89.88%	18.22%	4.93	69.21%	26.12%	2.65

Appendix Table A2. Performance of Daily News Tone Portfolios Based on Heterogeneous News Types

The table reports the performance of equal-weighted (EW) and value-weighted (VW) long-short (L-S) portfolios sorted on news tones and their long (L) and short (S) legs, based on three different types of news. News types include firm announcements (48.27% of all news articles), operation news (21.74% of all news articles) and equity news (17.72% of all news articles), which together constitute 87.73% of all news articles. The news types are categorized by SmarTag database. The decile portfolios are built based on the ensemble model of various LLMs. In Panel A, "Ret" and "*t*-Stat" stand for each portfolio's annualized return and *t*-Statistics. In Panel B, "Alpha" and "*t*-Stat" stand for each portfolio's annualized CH4-adjusted return and *t*-Statistics.

Panel A. Raw Returns (Annualized) for Long and Short Portfolios

News Type		EW		VW	
		Ret	*t*-Stat	Ret	*t*-Stat
Firm Announcements	L	44.32%	3.59	21.98%	1.61
	S	-56.71%	-3.89	-68.18%	-4.77
	L-S	101.04%	7.90	90.16%	5.81
Operation News	L	20.06%	1.73	15.51%	1.25
	S	-29.37%	-2.39	-29.20%	-2.26
	L-S	49.43%	4.49	44.71%	3.02
Equity News	L	54.47%	4.20	49.02%	3.05
	S	-34.77%	-2.54	-50.03%	-3.43
	L-S	89.24%	6.43	99.05%	5.23

Panel B. CH4-adjusted Returns (Annualized) for Long and Short Portfolios

News Type		EW		VW	
		Alpha	*t*-Stat	Alpha	*t*-Stat
Firm Announcements	L	35.33%	4.44	16.43%	1.64
	S	-67.81%	-7.38	-75.65%	-6.94
	L-S	103.13%	8.58	92.08%	6.15
Operation News	L	11.13%	1.68	9.77%	1.13
	S	-39.75%	-4.62	-37.56%	-3.67
	L-S	50.88%	4.75	47.33%	3.35
Equity News	L	44.94%	5.41	42.06%	3.31
	S	-46.99%	-4.72	-59.41%	-5.03
	L-S	91.93%	6.81	101.47%	5.51

Appendix Table A3. Sharpe Ratios of Portfolios Sorted by the Cross-Section of Return Predictions

The table reports the Sharpe Ratios of equal-weighted (EW) and value-weighted (VW) long-short (L-S) portfolios and their long (L) and short (S) legs. The portfolios are built based on BERT, FinBERT, RoBERTa, Baichuan, ChatGLM, and the ensemble model, respectively, using the cross-section of expected returns as sorting variables. "Ret", "Std", and "SR" stand for each portfolio's annualized return, standard deviation, and Sharpe Ratio, respectively

Model		EW			VW		
		Ret	Std	SR	Ret	Std	SR
BERT	L	27.11%	24.89%	1.09	19.39%	25.31%	0.77
	S	-29.38%	24.11%	-1.22	-16.13%	24.57%	-0.66
	L-S	56.49%	14.92%	3.79	35.52%	21.34%	1.66
FinBERT	L	35.05%	24.70%	1.42	20.00%	25.87%	0.77
	S	-38.16%	24.15%	-1.58	-25.64%	24.24%	-1.06
	L-S	73.21%	15.62%	4.69	45.64%	22.06%	2.07
RoBERTa	L	28.14%	25.09%	1.12	20.47%	25.71%	0.80
	S	-38.21%	24.05%	-1.59	-17.54%	24.39%	-0.72
	L-S	66.35%	15.37%	4.32	38.01%	21.79%	1.74
Baichuan	L	37.49%	23.93%	1.57	14.91%	24.97%	0.60
	S	-38.68%	24.55%	-1.58	-26.21%	24.99%	-1.05
	L-S	76.17%	16.04%	4.75	41.12%	22.30%	1.84
ChatGLM	L	33.70%	24.07%	1.40	15.75%	25.92%	0.61
	S	-33.26%	24.50%	-1.36	-19.76%	24.13%	-0.82
	L-S	66.96%	15.48%	4.33	35.51%	21.39%	1.66
Ensemble	L	38.65%	24.64%	1.57	14.66%	26.05%	0.56
	S	-40.88%	24.83%	-1.65	-26.85%	24.69%	-1.09
	L-S	79.53%	16.47%	4.83	41.51%	23.16%	1.79

Appendix Table A4. Performance of Portfolios With Stocks Trending Up or Down

The table reports the performance of decile portfolios sorted on daily news tones or return predictions, with stocks trending up or down. We first sort stocks into two subgroups based on past week returns. “Up” (or “Down”) indicates the group with positive (or negative) past week returns. Then, we sort on daily news tones or predicted returns based on the ensemble model of various LLMs, and form equal-weighted decile portfolios. Decile portfolios with the lowest (or highest) news tones or return predictions are denoted by Low (L) (or High (H)). H-L denotes the long-short portfolios. We report annualized return for each portfolio.

	News Tones		Predicted Returns	
	Up	Down	Up	Down
High (H)	33.17%	54.49%	13.67%	77.65%
Low (L)	-74.48%	-33.01%	-63.60%	-40.68%
H-L	106.69%	89.56%	77.97%	115.07%

Appendix Table A5. Long-Run Return Prediction Using News Tones

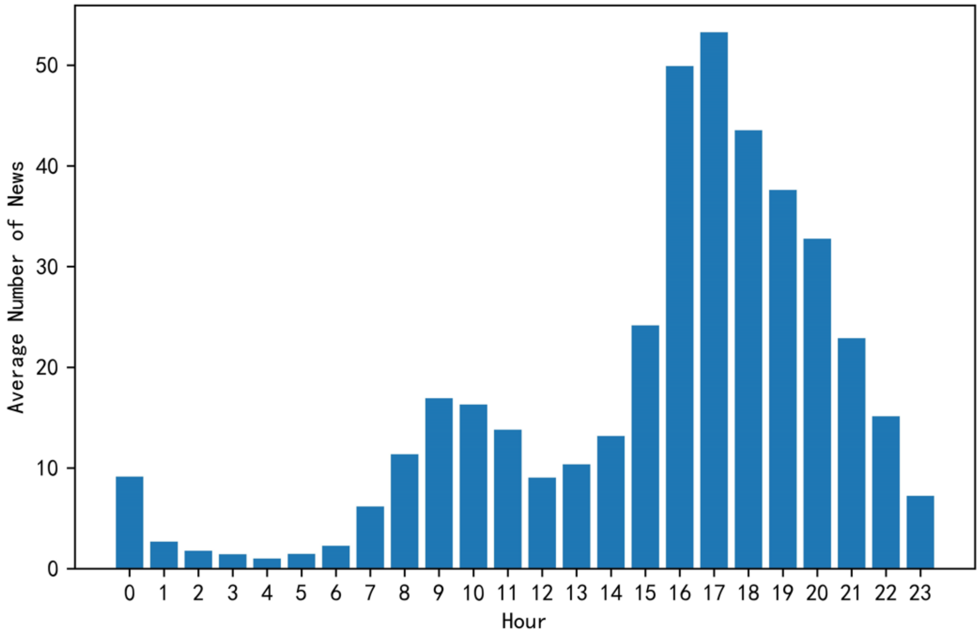
This table reports the long-run return prediction results using Fama-Macbeth regression. The dependent variable is the long-run return. Future horizon varies from 1 week to 8 weeks. Next *k*-week return denotes the average of the five daily returns in week *k*. The key independent variable, *Tone*, is the news tone extracted by the ensemble model. We include control variables in the regression, including previous open-to-open return (*LRet*), previous week open-to-open return (*Lwret*), previous month open-to-open return (*Lmret*), size (*Lsize*), EP-ratio (*Lep*) and turnover (*Lturn*). Newey-West adjusted standard errors are calculated using six lags. For simplicity, coefficients for control variables are not exhibited.

Dep. Var	Next *k*-week return	
	Tone Coef	*t*-Stat
k=1	0.0078	6.88
k=2	0.0008	0.78
k=3	0.0005	0.46
k=4	0.0007	0.70
k=5	0.0006	0.71
k=6	0.0016	1.83
k=7	0.0011	1.29
k=8	0.0013	1.54

Appendix 4

Appendix Figure A1. China News Count: Intraday Pattern

This figure plots the average number of news articles per hour (24-hour local time) in China, from January 2008 to December 2023.



Appendix Figure A2. Chinese Word Clouds of Long and Short Portfolios

This figure plots the word clouds of long and short portfolios using the BOW model.

Panel A. Long Portfolios' Word Cloud

表决
发展
提问
独立
投资建议
激励持续
高作为
销售
提升
实现
服务
同比增长
是
累计
净增长
创新
生产
建设
新
管理
变动
设备
保证收入
率
占
大
关注
投资
营业收入
表示
推动
公开
不构成
制造
同比下降
回答
合作
发布
经营
进行
摘录
提供
完成
工作
产能
上升
平安
开展
达
有
会
安
加工
打造
应用
升级
健康
保障
安全
增加
需求
审议
推进
相关

Panel B. Short Portfolios' Word Cloud

重组
投资
披露
重大
截至
收到
公开发行
简称
担保
亏损
控股股东
申请
辞职
控股
相关
关注
存在
实施
影响
经营
显示
说明
预告
预计
发行
减持
表示
拟
信息披露
警示
发布
营业收入
发布公告称公布
审计
重整
以下简称
购买进行
持有
是否
退市
工作
导致
开市
履行
支付
终止
冻结
同比下降
回复
审议
波动
收购
要求
深
净
回购
发展