Large Language Models and Return Prediction in China

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

Lin Tan, Huihang Wu and Xiaoyan Zhang*

Abstract

We examine whether large language models (LLMs) can extract contextualized representation of Chinese public news articles to predict stock returns. Based on representativeness and influences, we consider seven LLMs: BERT, RoBERTa, FinBERT, Baichuan, ChatGLM, InternLM, and their ensemble model. We show that news tones and return forecasts extracted by LLMs from Chinese news significantly predict future returns. The value-weighted long-minus-short portfolios yield annualized returns between 35% and 67%, depending on the model. Building on the return predictive power of LLM signals, we further investigate its implications for information efficiency. The LLM signals contain firm fundamental information, and it takes two days for LLM signals to be incorporated into stock prices. The predictive power of the LLM signals is stronger for firms with more information frictions, more retail holdings and for more complex news. Interestingly, many investors trade in opposite directions of LLM signals upon news releases, and can benefit from the LLM signals. These findings suggest LLMs can be helpful in processing public news, and thus contribute to overall market efficiency.

Keywords: return prediction, news articles, large language models, information efficiency, Chinese stock market.
JEL Codes: C52, C5, G1, G14.

* 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 Kaiji Chen, Lei Chen, Lin William Cong, Byoung-Hyoun Hwang, Fuwei Jiang, Raymond Kan, Xing Liu, Xiumin Martin, Jun Tu, Xiaolu Wang, Dacheng Xiu, Nianhang Xu, Bernard Yeung, Xintong Zhan, Xingjian Zheng, Guofu Zhou, conference participants at 2024 ABFER-JFDS Conference on AI and Fintech, 2024 China Fintech Research Conference (CFTRC), 2024 Summer Institute in Finance (SIF) Annual Conference, and seminar participants at Sun Yat-Sen University, Tsinghua University and 2024 Summer Institute in Digital Finance (SIDF) 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 public news articles to predict stock returns. Based on representativeness and influences, we consider seven LLMs: BERT, RoBERTa, FinBERT, Baichuan, ChatGLM, InternLM, and their ensemble model. We show that news tones and return forecasts extracted by LLMs from Chinese news significantly predict future returns. The value-weighted long-minus-short portfolios yield annualized returns between 35% and 67%, depending on the model. Building on the return predictive power of LLM signals, we further investigate its implications for information efficiency. The LLM signals contain firm fundamental information, and it takes two days for LLM signals to be incorporated into stock prices. The predictive power of the LLM signals is stronger for firms with more information frictions, more retail holdings and for more complex news. Interestingly, many investors trade in opposite directions of LLM signals upon news releases, and can benefit from the LLM signals. These findings suggest LLMs can be helpful in processing public news, and thus contribute to overall market efficiency.

Keywords: return prediction, news articles, large language models, information efficiency, Chinese stock market.
JEL Codes: C52, C5, G1, G14.

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 in U.S. and other 16 international stock markets in 13 different languages.[1]

In this study, we examine whether LLMs can extract signals from Chinese news texts to forecast stock returns in Chinese A-share market, the second largest market in the world. Chinese, as a language, is materially different from English and might pose a serious challenge for signal extraction. A major difficulty in processing Chinese is to handle the ambiguities of word segmentation and accurately identify word boundaries based on the surrounding linguistic context. For instance, the sentence "上市公司停牌原因说明会延期" can be interpreted in two different ways depending on word segmentation. The first version, "上市公司/停牌/原因/说明会/延期" means "the briefing on the reasons for the listed company's trading suspension is postponed"; and the second version "上市公司/停牌原因/说明/会延期" means "the listed company announces

[1] Their study includes English, Spanish, Italian, Chinese, and other languages. They study Chinese news for the HK market, rather than in the Chinese A-share market.

that the reasons for its trading suspension are likely to persist". Most previous studies, which rely largely on simplistic natural language processing (NLP) methods, focus on phonetics languages. Chinese, however, an ideographic character combination conveys meanings in a more compact and context-dependent way. Simple NLP methods, such as dictionary and bag-of-word (BOW), rely on predefined dictionaries and completely ignore word boundaries, word order and cross-word relations, constraining their capabilities in effectively processing Chinese.

LLM has the property of learning contextualized representations directly from character sequences, through the self-attention mechanism of the transformer architecture, and might help to process Chinese. Since it learns in an implicit, flexible and data-driven way, it avoids relying on explicit rules in word segmentation. The powerful LLMs clearly provide an exciting opportunity for efficiently processing various languages, while their application for Chinese capital markets remains unexamined. In this study, we focus on two research questions. First, can news tones and text features estimated using LLMs predict stock returns in China? Second, whether signals from LLMs help the price discovery process and contribute to market efficiency.

Since this study is about Chinese news and Chinese stock returns, 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. Combining the large quantity of retail traders and their low financial literacy, it is reasonable to expect that

these retail investors have difficulty in processing public information and trading on public news. Consequently, there might exist substantial under-processed information in public news for LLMs to extract. Second, with Chinese capital market's relatively short history, the overall information efficiency is still low, but it has been gradually improving over the past decades, as documented in Carpenter et al. (2021). Thus, it is reasonable to expect that, by rapidly analyzing and disseminating news, LLMs can play a significant role in accelerating this process and fastening price discovery.

We obtain news data from the ChinaScope SmarTag, which has a wide coverage of news sources and provides full content of every news article. Overall, the dataset contains 28 million news articles between January 2008 and December 2023, all written in Chinese. The text sources from 8,732 sites, including financial medias, government websites, and various entity's public portals in WeChat. The news articles cover 5,255 stocks in total, that is, 100% of the A-share stocks.

We employ the following criteria to select LLMs for our analysis: 1) trained on Chinese texts, to ensure comprehension of Chinese news; 2) open-sourced, including public model weights (the learned parameters representing the model's acquired knowledge) and technical documentation (comprehensive records on the model's architecture, training methodology, and implementation), allowing us to construct predictive signals; 3) influential. We assess a model's influence along two dimensions: academic influence (whether commonly adopted by existing studies) and industry influence (whether officially approved by the Cyberspace Administration of China for business use). Six individual LLMs satisfy these criteria: BERT, FinBERT, RoBERTa, Baichuan, ChatGLM, and InternLM. Specifically, BERT introduced by Devlin et al. (2018) forms the foundation of modern NLP and pioneers the transformer architecture. FinBERT and RoBERTa, both BERT

variants, offer enhanced understanding of financial contexts and improved training stability. In contrast, ChatGLM (Zeng et al., 2023), Baichuan (Yang et al., 2023), and InternLM (Cai et al., 2024) employ distinct architectures from BERT and represent the most recent developments of China's LLMs at the time of writing. Finally, to synthesize the information and provide a unified interpretation, we construct an ensemble model that aggregates the signals from individual LLMs.

To answer the first research question of return prediction, we proceed in three steps. First, we use pre-trained LLMs to convert news texts into numerical vector representations. These vectors are high-dimensional arrays of numbers that proxy for the semantic meaning of the context. Second, we use these vector representations to form news tone and construct return forecast. Third, we use the news tone and return forecast signals from previous day to form long-short portfolios and hold it for one day. If we use news tones, the annualized value-weighted returns for the long-short strategy range between 35.09% and 66.54%, all with significant t-statistics. If we use return forecasts, the returns range between 33.65% and 47.52%, and again all significant statistically. After risk adjustments, the abnormal returns are still around 38.14% to 69.90% (35.88% to 51.75%) when using news tones (return forecasts). That is, the LLMs can extract valuable information from Chinese public news and predict future stock returns in China.

Regarding the second research question on whether LLMs are helpful in incorporating public news information into prices, we proceed in four steps. First, we find that LLM signals can predict future earnings surprises, which suggest that the LLM signals contain information regarding firm fundamentals. Second, we find that the predictive power of LLM signals is higher when firms' information environments are less transparent, when retail investor holdings are higher and when

news are more complex. Third, we show the predictive power of LLM signals remains positive over two days after news releases, but diminishes after the third day. This dynamic indicates that the news information captured by LLMs is absorbed into prices within two days. Finally, we find different investors load their trades differently on the LLM signals, and some of them can benefit from LLM signals. These findings suggest LLMs can be helpful in processing public news, and thus contribute to overall market efficiency. We conduct a battery of robustness checks and find our empirical results are generally robust for different settings.

Our study naturally connects to the literature of textual analysis. Earlier 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) already demonstrate return prediction using former textual techniques such as word count and topic modelling. LLMs offer the most advanced technique, and multiple recent papers examine how LLMs help to predict returns. For instance, Chen, Kelly, and Xiu (2023) shows various LLMs can be used to process news articles and forecast stock returns in multiple markets; Lopez-Lira and Tang (2023) use ChatGPT to interpret news headlines and predict stock returns; Kim and Nikolaev (2023) use BERT to capture profitability from annual reports and examine asset pricing factor models. Together, this emerging literature suggests LLMs excel at extracting informative signals about asset prices and impact information dissemination and market efficiency.[2]

[2] Text data is also used in predicting market volatility as in Manela and Moreira (2017), modelling business cycles in Bybee et al. (2023a), and proxying for latent ICAPM state variables in Bybee et al. (2023b). In addition, 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.

There are few studies utilizing cutting-edge LLMs to extract news information and predict future stock returns in China. 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. Recently, Zhou, Fan and Xue (2024) propose a new machine learning model for stock return prediction in China, but they only use BERT, as a comparison benchmark, and rely on a single source of news media. Similarly, Jiang et al. (2024) employ their refined FinBERT for sentiment analysis in China, without considering the universe of domestically developed well-performing Chinese LLMs. 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 LLMs. Section 3 explains the empirical method. Section 4 presents the main results of predicting future returns using LLMs. We discuss the relation between LLMs and information efficiency in Section 5. Section 6 provides robustness results. Section 7 concludes.

2. Data and LLMs

2.1 Data

We obtain Chinese news text from ChinaScope SmarTag, a database provided by Mikuang Technology. This dataset offers four key advantages. First, it has an extensive coverage of real-time online public news, spanning financial medias, government sources, and WeChat official accounts, across 966 registered internet news providers. Considering there are 1,358 registered

internet news providers in China, the coverage of the dataset seems adequate.[3] Second, the news that the dataset covers are free and is accessible to all audience, which allows for significant impact. Third, compared to other domestic online news datasets (such as WIND, CNRDS and EastMoney), which provides only the title, our dataset provides the full content of news articles and exact reported timestamp, which allows for more in-depth analysis. Fourth, while many studies analyzing text data in China rely on newspaper datasets, such as Qin et al. (2018) who use the WiseNews covering 117 newspapers, our dataset provide real time news, which is more timely than print media.

We obtain over 28 million news articles from January 2008 to December 2023. Following Chen, Kelly, and Xiu (2023), we apply four filters to the news data. First, only articles that can be mapped to stocks are retained. Macroeconomic, industry, and entertainment news that are difficult to associate directly with stocks are removed. Second, to ensure each article is related to a single stock, we remove articles tagged with more than one stock. When an article mentions multiple stocks, we cannot label its representation with a single stock return, which creates accuracy issue for the subsequent economic modelling. Third, we only retain news related to Chinese A-share stock market. News on stocks from the U.S., Hong Kong, or other countries is removed. Fourth, we require stocks have available returns data, and remove unlisted and delisted firms.

As shown in Panel A of Table 1, the initial dataset contained 28,259,596 raw articles. A total of 8,372,112 articles are tagged with a single stock. After removing 6,138,364 news items on B/H-

[3] The number of internet news media is sourced from the 2021 statistics from the Cyberspace Administration of China's "List of Internet News Information Sources": https://www.gov.cn/xinwen/2021-10/20/content_5643834.htm.

share and other international stocks, 2,233,748 articles referencing A-shares remain. Finally, after removing 40,377 articles without returns, the remaining 2,193,371 articles are the main sample news for the following empirical analysis. Each year, a typical firm has on average 59.09 articles, which is comparable to a typical firm in the U.S., as in Chen, Kelly, and Xiu (2023).

We merge the news data with price, trading and accounting data from WIND. Chinese stock market opens at 9:30 am, and closes at 3:00 pm. Since China has a "T+1" trading rule in place, here we focus on daily returns rather than intra-day returns. The general norm for daily return is to use close-to-close return. However, as shown in Appendix Figure A1, the majority of news in China is released after the market close. Therefore, we compute open-to-open daily returns instead, to incorporate the overnight information into trading by forming portfolios at the next market open. Following Ke et al. (2019) and Chen, Kelly, and Xiu (2023), we compute daily return using market open prices, and daily return is defined as $ret_{i,t} = \frac{OpenPrc_{i,t+1}}{OpenPrc_{i,t}} - 1$, where $OpenPrc_{i,t}$ is the price of stock *i* on day *t* at the market open, after adjusting for splits and dividends. Table 1 Panel B shows that the mean return is 0.06% (15% annualized) with a standard deviation of 3.85%.

We obtain accounting data items to compute firm-level characteristics. They are used as control variables in the cross-sectional return prediction specifications, sorting variables in the heterogeneity analysis, and proxies for firm fundamentals in the mechanism analysis. These items include 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), and turnover (daily share trading volume divided by tradable shares outstanding).

Table 1 Panel B shows that the mean size is around 11.78 billion CNY with the median being 3.74 billion CNY. The average EP ratio and turnover is 0.48% and 2.85%, respectively.

Finally, we obtain data on stock's retail ownership, state ownership and analysts' coverage from CSMAR and SunTime. We define retail ownership as the proportion of shares held by non-institutional investors. The identification of state ownership follows Leippold et al. (2021). Coverage of analysts that issue earnings forecasts follows the method in Chan and Hameed (2006).

2.2 LLMs

Since the release of ChatGPT in late 2022, there is a rapid proliferation of research efforts and institutional investments in developing LLMs worldwide. In China, the number of large AI models registered with the Cyberspace Administration of China is eight in August 2023 at the time of writing, and quickly reaches 117 in March 2024. This proliferation presents a challenge for selecting appropriate models to examine our research questions. The choice of LLMs can significantly impact the quality and reliability of results. Thus, it is crucial for us to establish a set of selection criteria that aligns with our research objectives and practical constraints.

We select LLMs according to three criteria. First, models need to be trained on Chinese texts, to ensure comprehension of Chinese news articles.[4] Second, models need to be fully open-sourced to enable model fine-tuning for constructing predictive signals. We require modes to have complete weights and detailed technical documentation of architecture, training methodology, and implementation.[5] We don't consider closed-sourced LLMs in the main results, because these

[4] Due to our criterion of selecting only Chinese-specific models trained on Chinese texts, we do not include English-based models such as OPT (Zhang et al., 2022) and Llama (Touvron et al., 2022), despite their strong capabilities.
[5] The weights refer to the learnable parameters within the model that determine how input data is transformed into

models prevent us from inspecting the model structure and algorithmic details, and from maintaining the result reproducibility, due to its random model updates which might directly overwrite the previous series of models. Third, models need to be influential. We assess a model's influence in two dimensions: academic influence (whether commonly adopted by existing studies) and industry influence (whether officially approved by the Cyberspace Administration of China).[6] These three criteria together ensure the selected models' representativeness and influence.

Given the above criteria, we mainly study six individual LLMs and one ensemble model.[7] The first model is BERT. BERT is a pretrained language model originally introduced by Devlin et al. (2018). For our study, we adopt the version optimized for Chinese, that is, the bert-base-chinese model from Hugging Face. It is trained on Chinese text (including 1.1 billion words from Chinese Wikipedia) and becomes fully open-sourced since November 2018. It has a high influence in the academic: its original BERT version in English is adopted by a range of studies including Chen, Kelly and Xiu (2023) and Kim and Nikolaev (2023), and its downloads of 55 million times is the highest among all LLMs from Hugging Face. Technically, it innovatively undertakes two unsupervised objectives: masked language modelling and next-sentence prediction.[8] Such training

output predictions. These weights are determined during the pre-training process. A model with public and complete weights allows us to inspect and further fine-tune the model.

[6] In August 2023, the Cyberspace Administration of China, together with other government agencies, issue the "Interim Measures for the Management of Generative Artificial Intelligence Services". This regulation is the first specifically addressing generative AI models, stipulating Chinese LLM providers must undergo security assessments and register algorithms before offering commercial generative AI services. Approved models are perceived to meet government standard for capability and security.

[7] We compare these models with ChatGPT in our robustness check.

[8] The masked language modeling 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. The next-sentence prediction involves training to determine whether two given text segments maintain sequential coherence with respect to their original contexts.

process teaches the model word relationships, thereby equipping it with knowledge that can be transferred to downstream tasks through fine-tuning. Its strength lies in its representativeness, as it acts as a foundation benchmark for all LLMs. Its limitations include training with a relatively small batch size (further refined by RoBERTa) and few financial contexts (improved by FinBERT).

The second model is RoBERTa. RoBERTa is an optimized replication of pre-training BERT. We use the Chinese adaptation of the original RoBERTa, a large version of XLM-RoBERTa proposed by Conneau et al. (2019).[9] It is trained on Chinese and becomes fully open sourced since December 2019. The RoBERTa series of models has a high academic influence with the examination of Chen, Kelly and Xiu (2023), with over 36 million downloads from Hugging Face. It builds on key ideas from the BERT, but modifies the hyper-parameters, 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 increase training stability and enhance the performance compared with BERT across NLP benchmarks. However, the potential weakness of BERT of lacking training on financial data remains with RoBERTa.

The third model is FinBERT. FinBERT raised by Yang et al. (2020) is a financial domain-specific model fine-tuned on a large scale of financial texts based on BERT pre-trained parameters. We use an open-sourced Chinese FinBERT model, Valuesimplex FinBERT.[10] Public since October 2020, this is the first open-source Chinese model pre-trained on Chinese financial texts containing 30 billion tokens from financial news, analyst reports, company announcements, and financial

[9] XLM-RoBERTa is an adaptation of RoBERTa pre-trained on 2.5 terabytes of filtered CommonCrawl data learning useful representations across multiple languages. Specifically, we use xlm-roberta-large from Hugging Face.
[10] Model sources from https://github.com/valuesimplex/FinBERT.

encyclopedia entries. Its original version has a high academic influence with the test of Chen, Kelly and Xiu (2023), with around 2 million downloads from Hugging Face. Previous experiments show FinBERT has strong performance on downstream financial NLP tasks, such as financial sentiment analysis and relation extraction. Its potential weakness is that the number of parameters inherited from the original BERT may restrict its modeling capacity for highly complex and long contexts.

The fourth model is the ChatGLM model. Designed by Zeng et al. (2023), it is an open bilingual language model with the Du et al.'s (2022) general language model (GLM) architecture. We adopt the most recent open-source version, the ChatGLM3-6B.[11] It is among the first batch of models approved by the Cyberspace Administration in August 2023. Since March 2023, the initial open-source version, ChatGLM3-6B has over 10 million downloads on Hugging Face, demonstrating a high level of interest and usage. 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 Chinese-English training supplemented by supervised fine-tuning, self-supervised feedback, human-in-the-loop reinforcement learning, and other techniques. It can also form a unified token set with over 150 thousand vocabularies. The potential weakness of ChatGLM, however, includes the general-purpose pretraining which may limit its capacity to capture financial predictive signals.

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 trained on 2.6 trillion tokens.[12] The Baichuan series

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

has over 120 thousand downloads on Hugging Face since the initial open-source in June 2023, demonstrating its popularity. It is among the first batch of approved models by the Cyberspace Administration of China in August 2023. Unlike the GLM architecture used by ChatGLM, Baichuan-2 uses a model architecture consistent with LLAMA, that is, transformer's decoder-only architecture. To better handle Chinese text, Baichuan uses new techniques such as the Byte Pair Encoding (BPE) tokenizer, applying no normalization to input text, and incorporating additional whitespace tokens to handle long phrases. With the new techniques, Baichuan's advantages include matching or exceeding other open-source models of comparable sizes on public benchmarks, while the lack of training on financial data remains a potential drawback.

The sixth model is InternLM. It is mainly developed by the Shanghai AI Laboratory in 2023. We adopt the most recent open-source version, the InternLM-7B.[13] Since its initial release in June 2023, InternLM attracts interest of over 370 thousand downloads on Hugging Face. A potential strength of InternLM is its comprehensive data preparation, including pre-training data and domain-specific enhancement data. The limitation of the model may lie in its potential inability to fully comprehend the nuances of financial contexts.

With the six individual LLMs, we create an ensemble model to extract common signals, synthesize distinct predictors, and provide a unified economic interpretation. As suggested by Dietterich (2000), the ensemble method improves the overall accuracy and robustness. Statistically, ensemble may improve the prediction accuracy by reducing the overall variance and canceling out

[13] Specifically, we use the model internlm2-base-7b from Hugging Face.

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 calculate the arithmetic mean of the signals from six individual LLMs for each firm-day observation. Simple averaging offers an easy, less capacity-consuming, and interpretable methodology. 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.

3. Extracting Signals from Text Data Using LLMs

3.1 Model Set-up

Following Ke et al. (2019), our study extracts two key signals using financial news text: news tone and return forecast. The first signal, news tone, provides estimate for the possibility of a news articles conveying positive information. The second signal, return forecast, provides estimate for the news' associated future stock return.

For forming news tones from news articles, we estimate a logistic model:

$$E\left(y_{i,t+1}\middle|x_{i,t}\right) = \frac{e^{x_{i,t}'\beta}}{1+e^{x_{i,t}'\beta}} . \quad (1)$$

The logistic function is specifically designed to map the independent variable to a range between 0 and 1, thus standardizing the tone quantification. The dependent variable, $y_{i,t+1}$, is a labelled dummy variable for stock *i* on day *t+1*. It takes the value of 1 (or 0), that is, $y_{i,t+1} = 1$ (or $y_{i,t+1} =$

0), when the next-day return $r_{i,t+1}$ is positive (negative) for stock *i* on day *t+1* in the training sample. For each news released on day *t* corresponding to stock *i*, $x_{i,t}$ is the LLMs' article-level representation, a high-dimensional numerical vector.

This article-level representation vector is derived through the following steps: news tokenization, transformer architecture, pre-training, fine-tuning, and news embedding.[14] The news embedding step, given a tokenized article, maps each token to a high-dimensional embedding vector using the model's pre-trained embedding matrix.[15] Then, following Chen, Kelly, and Xiu (2023), we take the average of all token-level embedding vectors, which gives us the article-level representation, $x_{i,t}$. It can summarize full article content and represent the overall semantic information. Importantly, its dimensionality varies according to the specific architecture of LLMs. For BERT model, $x_{i,t}$ is a 1×768 vector; for RoBERTa and FinBERT, $x_{i,t}$ is a $1 \times 1{,}024$ vector; for ChatGLM, Baichuan and InternLM, $x_{i,t}$ is a $1 \times 4{,}096$ vector.

For instance, Figure 1 provides an example of deriving an article-level representation using the Baichuan model. Given the example news mentioning the Bank of Ping'An, Baichuan first tokenizes it into 71 tokens. For each token, Baichuan's pre-trained embedding matrix transforms the token into a vector of 4,096-dimension. The article-level representation, $X_{i,t}$, is then derived

[14] Details on the first four steps are in Appendix 1.

[15] The total number of tokens mappable per article is limited for each model. BERT-based models can process up to 512 tokens, while ChatGLM, Baichuan and InternLM can handle around 8k, 4k, and 200k tokens. Appendix Table A1 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, 161 for ChatGLM, and 152 for InternLM. Given the token number's distribution, 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 processed with all tokens embedded for BERT-based (Baichuan, ChatGLM and InternLM) models.

by averaging the 71 embedding vectors, which is a high-dimensional vector with a length of 4,096.

Similarly, for constructing the second signal, the return forecasts, we estimate a linear model:

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

where $r_{i,t+1}$ is the labelled continuous variable of the next-day return for stock *i* on day *t+1* in the training sample. $x_{i,t}$ is the same LLMs' article-level representation used in Equation (1).

3.2 Parameter Estimations

To estimate the parameter β and θ in Equation (1) and (2), we take the following two steps: choosing training sample, and minimizing error loss function for the trained observations. After obtaining parameter estimates, we apply to the testing sample, and calculate the out-of-sample news tones and return forecasts signals, which we further use for return prediction.

Our training sample period is from January 2008 to December 2018, while our testing sample is from January 2019 to December 2023. We utilize an expanding-window approach for model training. In the first iteration, we use data from 2008-2018 to get the first-round estimation of β and θ. The year 2019 is reserved for out-of-sample testing. In the second iteration, the training window expands to 2008-2019, 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.[16]

In training the model, we follow Chen, Kelly and Xiu (2023) and estimate parameters β by minimizing a standard objective function, the cross-entropy loss function with L2 penalty:

$$LossCE(\hat{\beta}) = -\frac{1}{N}\sum_{i=1}^{N}[y_{i,t+1}\, log(\widehat{y_{i,t+1}}) + (1 - y_{i,t+1}) \log(1 - \widehat{y_{i,t+1}})] + \frac{\alpha}{2N}\sum_{j=1}^{M}\hat{\beta}_j^2, \quad (3)$$

where $\hat{\beta}$ is the estimate for the true parameter β, $\widehat{y_{i,t+1}} \equiv \frac{e^{x'_{i,t}\hat{\beta}}}{1+e^{x'_{i,t}\hat{\beta}}}$ is the estimated dependent variable with parameter estimate $\hat{\beta}$, N is the total number of observations in the training dataset, M is the total dimensionality of the vector $\hat{\beta}$ (each dimension is denoted by *j*), and α is a tuning parameter. The benefit of using the L2 penalty term, $\frac{\alpha}{2N}\sum_{j=1}^{M}\hat{\beta}_j^2$, is to address overfitting due to the high dimensionality of the $x_{i,t}$. It is a form of regularization that adds the sum of squared coefficients multiplied by a penalty parameter, α, which is an optimized hyperparameter that controls the strength of the penalty.[17] This regularization helps keep the coefficients small, which can lead to a simpler model that is less likely to overfit the training data. The optimal sample estimate, $\hat{\beta}^*$, minimizes Equation (3). With the optimal parameter estimate $\hat{\beta}^*$ which we obtain from the training sample, we apply it to every $x_{i,t}$ in the testing sample, and we term the out-of-sample $\widehat{y_{i,t+1}}^* \equiv \frac{e^{x'_{i,t}\hat{\beta}^*}}{1+e^{x'_{i,t}\hat{\beta}^*}}$ the estimated news tone, *Tone*. The variable *Tone* ranges between 0 and

[16] Some may be concerned about our LLMs learning from future news data during pre-training. This is not a valid concern for two reasons. First, the basic model we use, Google's BERT, is 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 public models 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 is "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.

[17] We determine the optimal α using 5-fold cross-validation with a grid search over the log range of e^{-5} to e^{5}. 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 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.

1, with value closer to one meaning the stock more likely associated with positive future return.

Similarly, we follow Chen, Kelly and Xiu (2023) and estimate parameters θ by minimizing the mean squared error (MSE) loss function with L2 penalty:

$$LossMSE(\hat{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\left(r_{i,t+1} - \widehat{r_{i,t+1}}\right)^2 + \frac{\alpha}{2N}\sum_{j=1}^{M}\hat{\theta}_j^2 \ , \quad (4)$$

where $\hat{\theta}$ is the parameter estimate for θ, $\widehat{r_{i,t+1}} \equiv x_{i,t}'\hat{\theta}$ is the estimated dependent variable with parameter estimate $\hat{\theta}$, N is the total number of observations in the training dataset, M is the total dimensionality of the vector $\hat{\theta}$ (each dimension is denoted by *j*), and α is a tuning parameter. The optimal sample estimate, $\hat{\theta}^*$, minimizes Equation (4). With the optimal parameter estimate $\hat{\theta}^*$ which we obtain from the training sample, we apply it to every $x_{i,t}$ in the testing sample, and we term the out-of-sample $\widehat{r_{i,t+1}}^* \equiv x_{i,t}'\hat{\theta}^*$ the estimated return forecast, *RetForecast*.

3.3 Model Fits

We examine the fits of our models in sample and out-of-sample, and present these results in Table 2. For in sample estimation, we obtain the estimated parameters $\hat{\beta}^*$ and $\hat{\theta}^*$, then compute the *Tone* and *RetForecast* signals during training sample. Table 2 Panel A presents the in-sample distribution for these *signals*. We expect these signals to exhibit similar distributions across models if they well-fit the training data. Panel A reveals that all models produce same negative mean news tone signals of -0.0182 (except InternLM at -0.0181), with ChatGLM and Baichuan exhibiting the highest standard deviation. Similarly, all models generate same mean return forecasts of -0.0001, indicating consistent model fitting during the training period. For Panel B, we provide the basic distributions for the *Tone* and *RetForecast* signal in the testing sample. We expect variation in model performance when applied to new data in the testing period. We find BERT generates the most negative mean news tone (-0.0772), while Baichuan produces the least negative mean (-

0.0577). In terms of return forecasts, BERT-based models show slightly positive means (0.0001-0.0002), while newer LLMs (Baichuan, ChatGLM, InternLM) generate lower means near zero. For both signals, newer LLMs demonstrate higher volatility in their predictions compared to the BERT-based models.

For out-of-sample fit, we compute tone accuracy and cross-sectional correlation. Regarding tone accuracy, we first obtain the out-of-sample *Tone*, i.e., $\hat{y}^*$ derived in Section 3.2. Then, a true positive (true negative) prediction, *TP* (*TN*), occurs when $\hat{y}^* > 0.5$ (< 0.5) aligns with a realized positive (non-positive) return, $y = 1$ (0), on the next day. False positives (FP) and false negatives (FN) represent complementary cases in which news tones $\hat{y}^* > 0.5$ (< 0.5) coincides with a realized non-positive (positive) return, $y = 0$ (1) on the next day. We calculate the out-of-sample tone accuracy as the proportion of correct predictions (*TP* and *TN*) in the testing sample: *Accuracy* = $\frac{TP + TN}{TP + TN + FP + FN}$. If the model performs well, we expect the out-of-sample tone accuracy to exceed 50% (the threshold of exceeding random guessing); the model performs better when the tone accuracy approaches to 100%. Table 2 Panel C presents the out-of-sample prediction accuracy for the LLMs. We find that the seven models consistently outperform random guessing (accuracy = 50%) over the sample period. Among all models under study, the FinBERT model delivers the highest overall accuracy of 52.77%, exceeding that of the ensemble model (52.74%), BERT (52.71%), RoBERTa (52.63%), Baichuan (52.37%), InternLM (52.08%), and the ChatGLM (51.93%). Previous findings of Chen, Kelly and Xiu (2023) in U.S. stock market report an accuracy of 53.41% for BERT, which is comparable to what we find here.

Regarding cross-sectional correlation, we obtain the out-of-sample *RetForecast*, i.e., $\hat{r}^*$ derived in Section 3.2, and calculate the time-series average of the cross-sectional correlations between each model's return forecasts $\hat{r}^*$ and the realized next-day returns r in the testing sample.

If the model performs well, we expect the cross-sectional correlation to be positive; the model performs better when the correlation approaches to 1. Table 2 Panel D reports the time-series average of the cross-sectional correlations between each model's return forecast and the realized next-day return in the testing sample. The correlations are positive for all models in testing years and generally higher than 1%. Baichuan and the ensemble model have the highest overall correlation of 1.99% and 1.95%, followed by the FinBERT model (1.94%). RoBERTa, InternLM, BERT, and ChatGLM's overall correlations are 1.80%, 1.73%, 1.62% and 1.52%, respectively. Previous findings in Chen, Kelly and Xiu (2023) find the overall correlation in U.S. stock market to be 1.62% for BERT, comparable to our statistics.

Overall, we find the model fit is consistent in sample and reasonable in testing-sample.

4. LLMs and Return Predictions

To measure whether the signals from LLMs can effectively predicting future returns, we adopt two approaches. First, we sort portfolio based on LLM signals and compute portfolio returns to assess the economic magnitude of return prediction. Second, we estimate Fama-MacBeth regressions to provide robust results after controlling for other firm-level characteristics.

4.1 Portfolio Returns

For the portfolio approach, we adopt an open-to-open trading timeline, which builds and rebalances positions at every market open, making use of news since the previous market open. Specifically, for all the news released from the open of trading day t till the open of the next trading day $t+1$, we compute their news tones and return forecasts from the LLMs, and construct portfolios at the open of day $t+1$ by sorting on stocks' news tones or return forecasts. For news released on

non-trading days, their news tones and returns forecasts are sorted at the first open of the subsequent trading day. The top decile portfolio includes firms with the most positive news tones or the highest return forecasts, and is denoted as the long-leg. The bottom decile portfolio includes firms with the most negative news tones or the lowest return forecasts, and is denoted as the short-leg. A zero-net investment long-minus-short strategy refers to taking long positions in the long-leg while taking short positions in the short-leg. Then, we hold the long, short, or long-minus-short positions for one day, until the open of the next trading day *t+2*, and rebalance at the open of day *t+2.* We examine the portfolios' value- and equal-weighted annualized returns. If a LLM performs well in extracting news for predicting returns, then we expect the long-minus-short strategies based on its signals to deliver significantly positive returns.

To further adjust for systematic risk exposures and calculate the alphas for the trading strategies, we adopt the CH4 factor model of Liu et al. (2019a), which is tailored to the Chinese stock market. We regress each portfolio's return on the contemporaneous returns of the market factor, size factor, value factor and turnover factor from the CH4 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. If a model performs well, then we expect the long-minus-short strategies to have significantly positive alphas.

Table 3 reports detailed returns on portfolios sorted by news tones. From Panel A where we examine raw returns, three key patterns emerge. First, across all LLMs and both weighting schemes, long legs consistently outperform short legs with the long minus short (L-S) strategies generating significantly positive annualized returns ranging from 35.09% (64.28%) to 66.54%

(88.52%) when VW (EW), demonstrating the economic value of the tone signals provided by LLMs. Second, among all LLMs, the Baichuan and ensemble model deliver the top two 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 88.52% with a *t*-Stat of 9.88 for the ensemble model, 84.40% with a *t*-Stat of 9.27 for Baichuan, 77.55% with a *t*-Stat of 8.99 for FinBERT, 74.26% with a *t*-Stat of 9.52 for RoBERTa, 72.71% with a *t*-Stat of 8.73 for InternLM, 65.26% with a *t*-Stat of 9.13 for BERT, 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 66.54% with a *t*-Stat of 5.57 for Baichuan, 63.64% with a *t*-Stat of 5.30 for the ensemble model, 54.88% with a *t*-Stat of 4.93 for RoBERTa, 51.08% with a *t*-Stat of 4.52 for FinBERT, 47.45% with a *t*-Stat of 4.60 for BERT, 37.37% with a *t*-Stat of 3.42 for ChatGLM, and 35.09% with a *t*-Stat of 3.24 for InternLM. Third, the EW long minus short returns exceed their VW counterparts, suggesting better prediction benefits in small-cap stocks. Taking the ensemble model for instance, its EW L-S return of 88.52% is higher than its VW counterparts of 63.64%.[18]

In Table 3 Panel B, we conduct risk-adjustment to the raw returns in Panel A using the CH4 factor model of Liu et al. (2019). Three key patterns are identified. First, all the models deliver significantly positive alphas. Second, the alphas become larger in magnitude when switching from

[18] The Sharpe ratio (SR) result of each model's long minus short strategy is provided in Appendix Table A2. Ranking the LLMs from delivering the highest SR to the lowest, the EW long minus short strategy has a SR of 4.97 for the ensemble model, 4.73 for RoBERTa, 4.62 for Baichuan, 4.47 for InternLM, 4.46 for FinBERT, 4.40 for BERT, and 3.97 for ChatGLM.

the VW to the EW scheme. Finally, when EW, 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 91.32% with a *t*-stat of 11.76, 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, InternLM's alpha being 75.41% with a *t*-stat of 10.09, 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.[19]

Similarly, as for our second signal, Table 4 presents detailed portfolio returns sorted by the return forecasts. 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, delivering significantly positive long minus short returns everywhere. Second, the ensemble model outperforms all other LLMs. Taking the VW scheme for instance, the VW L-S strategies generate significantly positive annualized returns of 47.52% for the ensemble model (*t*-Stat of 4.48), 45.64% for FinBERT (*t*-Stat of 4.45), 41.12% for Baichuan (*t*-Stat of 4.06), 38.01% for RoBERTa (*t*-Stat of 3.68), 35.52% for the BERT (*t*-Stat of 3.50), 35.51% for ChatGLM (*t*-Stat of 3.46), and 33.65% for InternLM (*t*-Stat of 3.18). Third, the EW method outperforms the VW, signaling higher return predictive power within small-cap stocks. Specifically, the EW L-S strategies yield significantly positive annualized returns of 80.20% for the ensemble model (*t*-Stat of 9.24), 76.17% for Baichuan (*t*-Stat of 9.36), 73.21% for FinBERT (*t*-Stat of 9.40), 68.95% for InternLM (*t*-Stat of

[19] Our results are robust after filtering out stocks suspended from trading, or hitting the price limits at the market open, when forming the news tone portfolios. The average annualized returns on the L-S, L and S portfolios and *t*-Stats are in Appendix Table A3.

8.33), 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). Finally, the results are similar and robust when returns are risk-adjusted using the CH4 model in Panel B. [20]

In Figure 2, we report each LLM's cumulative log returns for long minus short portfolios in the testing sample, sorted by news tones or return forecasts signals. Panel A and B reports value-weighted long minus short returns, and Panel C and D reports equal-weighted returns. Take Panel A when the value-weighted portfolios are sorted by the news tone signal for instance. We notice three key patterns. First, all LLMs' news tone signals cumulate increasingly positive log returns, ranging from 1.5 to 3.0, in the whole testing sample. Second, the Baichuan model, indicated by the highest red line, is the best-performing individual LLM. It is followed by RoBERTa, FinBERT, BERT, ChatGLM and InternLM. Third, all models substantially beat the market portfolio, which is the relatively flat black line. Panel B show similar value-weighted results when sorted by the return forecast signal. Panel C and D provide additional evidence that the equal-weighted portfolio returns are higher than the value-weighted, reaching a cumulative log return of around 4.0 for the best-performing ensemble model. It is thus evident that adopting LLMs to capture the textual implications of Chinese news has great value to add to the investment industry.

4.2 Fama-MacBeth Regressions

In this section, we estimate Fama-MacBeth (1973) regressions to examine the return predictive power of LLM signals. By this approach, we can control for stock-level characteristics

[20] In Appendix Table A4, we provide robustness results where we filter out stocks that are suspended from trading or hitting the price limits at the market open when forming the return forecast portfolios.

that might influence future returns. For each day *t*, we estimate the following cross-sectional specification:

$$r_{i,t+1} = a_{0,t+1} + a_{1,t+1}Tone_{i,t} + a'_{2,t+1}Controls_{i,t} + u_{i,t+1}\,, or,$$
$$r_{i,t+1} = a_{0,t+1} + a_{1,t+1}RetForecast_{i,t} + a'_{2,t+1}Controls_{i,t} + u_{i,t+1}\,. \quad (5)$$

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 returns forecasts can predict future stock price movements, we expect a significantly positive coefficient of a_1, the time-series average of $a_{1,t+1}$.

Table 5 demonstrates consistent return predictive power across LLMs. In Panel A, the positive and significant coefficients of the news tones indicate that higher tones are significantly associated with greater future returns. For instance, the coefficient for *Tone* is 0.0166 (1.66% daily) with a *t*-Stat of 10.86 for the ensemble model. Panel B shows similar results for the LLMs' return forecasts. For example, the coefficient for *RetForecast* is 0.4341 with a *t*-Stat of 11.67 for the ensemble model. Both panels suggest the significantly return predictive power of text-based signals, even after controlling for characteristics that might influence future returns.

To summarize, we provide robust evidence that the LLM signals fit in the testing sample, provide L-S strategies yielding significantly positive alphas, and cannot be overwritten by firm-level characteristics. This answers the first research question of whether the LLMs' signals can predict future stock returns.

5. LLM vs. Information Efficiency

The return predictive power of LLM signals raises important questions about market efficiency. The semi-strong form of an efficient market in Fama (1970) posits that all public information is immediately reflected in asset prices. However, a market that is not semi-strong efficient allows for certain public information taking time to be incorporated into prices. In this section, we investigate whether signals from LLMs help the information transmission process. In Section 5.1, we examine the information content of LLM signals, in terms of whether it is related to firm fundamentals. We examine the heterogeneity of firms to find in which situations the LLM signals have stronger predictive power in Section 5.2. In Section 5.3, we estimate the assimilation speed of the news information captured by LLM signals. We investigate how different investors react to the LLM signals in Section 5.4.

5.1 LLM Signals' Information Content

With the previous section answering the magnitude of prices incorporating LLM signals' contained information, we proceed to examine what specific information content that LLMs help discover from the public news. Tetlock et al. (2008) suggest that news conveys important fundamental information about firms' future performance. It is possible that the LLMs might help to contain firm fundamental information that is not yet fully incorporated into prices, which might be the reason for LLM signals' predictive power for return. Empirically, we examine whether LLM signals can predict future earnings surprises using panel regressions, as in Tetlock et al. (2008):

$$SUE_{i,t+1} = a_0 + a_1 Tone_{i,t} + a_2' Controls_{i,t} + u_{i,\,t+1}\,,$$
$$SUE_{i,t+1} = a_0 + a_1 RetForecast_{i,t} + a_2' Controls_{i,t} + u_{i,\,t+1}\,. \quad (6)$$

The dependent variable, $SUE_{i,t+1}$, is quarterly unexpected earnings for firm *i* on day *t+1*, with day

t+1 being the earnings announcement day. 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. The key independent variable is the news tone or return forecast signal from the ensemble model that occurs on the day *t*, before the earnings announcements. We include lagged controls similar to those in Table 5. Standard errors are double clustered at stock and calendar quarter level.

Table 6 presents the earnings prediction results. From Panel A, the news tone has a significantly positive coefficient of 4.63 with a *t*-stat of 8.65. That is, the news tone from day t significantly and positively predicts next-day earnings surprise. Similarly, in Panel B, the return forecast has a significantly positive coefficient of 135.31, meaning that the return forecast from LLMs can also positively and significantly predict future earnings surprise. These results suggest LLM signals contain information related for firm level fundamentals.[21]

5.2 LLM Signals' Predictive Power in the Cross Section

The effectiveness of LLMs in processing public news is not uniform across all firms and news. Our analysis is motivated by the economic intuition that while LLMs help extract hard-to-observe fundamental information from news (as shown in Section 5.1), the market's ability to efficiently incorporate such information, according to the assumptions in Efficient Market Hypothesis,

[21] Appendix Table A5 also shows that the positive return predictive power does not revert over longer horizons.

depends on information processing frictions, such as information acquisition costs, information environment, etc. Thus, the benefits of LLM-based news processing are likely to be more pronounced when these frictions are higher. More specifically, frictions are higher when the inattention of the prevalent retail investors creates more room for mispricing, when information environments are opaquer, and when the complexity and source of news makes it more challenging for market participants to quickly and accurately assess its implications.

At firm-level, we employ five important characteristics: 1) retail ownership; 2) size; 3) shorting volume; 4) state ownership; and 5) analyst coverage. Regarding retail ownership, the Chinese stock market, with retail investors contributing 80% of daily trading according to Jones et al. (2024), offers a unique setting for examining the role of retail in information processing efficiency. Liao et al. (2021) and Titman et al. (2022) suggest that retail exhibit inattention and constrained processing capabilities, which hinder them from interpreting firm fundamentals and acquiring timely information. Regarding size, Diamond and Verrecchia (1991) and Bhushan (1989) suggest firm size proxies for information environment, where smaller firms tend to have opaquer information environment with higher information asymmetry. Regarding shorting volume, short-selling constraints are stringent in China. Only a portion of stocks are eligible for margin trading and shorting, and among them, lendable supply matters. Saffi and Sigurdsson (2011) suggest that low shorting activity proxies for high shorting costs which allows overpricing to persist longer. This may impede the market's ability to correct mispricing even when fundamental information becomes available. Regarding state ownership, Jiang and Kim (2020) suggest state-owned firms have better corporate governance which helps reduce information asymmetry and build more

transparent information environment. Regarding analyst coverage, Chan and Hameed (2006) indicate stocks lacking analyst coverage potentially suffer from reduced fundamental information production and greater information asymmetry.

Empirically, we sort stocks into two subgroups based on the above proxies, then form decile portfolios within each subgroup. Results are in Table 7 Panel A. Take retail holding as an example. When sorted by news tones, for high retail ownership stocks, the long minus short portfolios yield annualized returns of 76.02% when VW (*t*-Stat 5.35), and 94.34% when EW (*t*-Stat 8.61). These are higher than for low retail ownership stocks, where the returns are 61.72% when VW (*t*-Stat 4.53) and 81.22% when EW (*t*-Stat 7.55). The differences of 14.30% when VW (76.02%-61.72%) and 13.12% when EW (94.34%-81.22%) are economically significant, given the average annualized return for the whole A-share market is around 10%. The difference is similar when sorted by return forecasts. Meanwhile, other subgroup results in Panel A show higher long minus short returns for firms smaller and less shorted, state-owned or covered by analysts.

At news-level, we adopt the following two proxies: 1) adoption of uncommon characters; and 2) central media source. For adoption of uncommon characters, following Wang et al. (2018), we compute the adoption of characters not included in the "List of Common Modern Chinese Characters". This list is published by the State Language Work Committee in 1988 and comprises 3,500 standardized characters. The presence of uncommon characters indicates specialized terminology, reflecting complexity of news. For central media source, we determine "central media" following the "List of Internet News Information Sources" published by the Cyberspace Administration. Such news is suggested to be less critical, accurate, comprehensive or timely than

those by market-oriented medias, according to You et al. (2018).

Empirically, we conduct similar subgroup sorting in Panel B. Regarding uncommon character adoption, when sorted by news tones, the long minus short strategy using news with uncommon characters generates a 98.30% annualized VW return (*t*-Stat of 6.53), substantially outperforming the 48.73% (*t*-Stat of 3.62) using simpler news. This pattern persists when using equal-weighted or sorting by return forecasts. Regarding media source, since central media news only account for 1.8% of all news, portfolio sorting becomes unfeasible, so we turn to a panel regression. In Panel C, we predict next-day stock return using the interaction between LLM signal and *CentralMedia.* *CentralMedia* is a dummy that takes 1 when the news is from central medias.[22] Panel C shows *Tone*CentralMedia* has a significantly positive coefficient of 0.0106 (*t*-Stat of 2.79), and *RetForecast*CentralMedia* has a significantly positive coefficient of 0.3019 (*t*-Stat of 2.83). That is, the return predictive power is stronger when the news is more complex and from central medias, consistent with our intuitions.[23]

5.3 Assimilation Speed of LLM-Extracted News Information

Real-world markets often exhibit inefficiencies due to various frictions and information processing constraints. Thus, understanding the speed at which LLM-extracted information is

[22] Control variables are similar to Table 5. We include both stock- and day-level fixed effects, and cluster the standard errors at day level following Petersen (2009).

[23] To further make use of news characteristics, we also consider news categories and other characteristics. In Appendix Table A6, we follow the dataset's categorization and separate three types of news: firm announcements, operation news and equity news. We find the annual returns on L-S portfolios sorted on firm announcements and equity news tones substantially higher (nearly twofold) than operation news, suggesting more undetected information in the former news types. Meanwhile, by computing the frequency of negation or numbers in each article, the L-S returns are higher when using news with higher negation ratio and wordier expression (with less illustrating numbers), suggesting the usefulness of LLMs in pricing information that needs more contexts and interpretation.

reflected in prices is crucial for several reasons. First, it provides insights into the persistence and duration of potential arbitrage opportunities. Second, it helps gauge the efficiency of Chinese stock market in processing public news. Lastly, it offers practical implications for the implementation of LLM strategies, particularly in determining optimal holding periods and trading frequencies.

Empirically, we test when initiating a strategy using already announced news could still earn positive returns. We compute the average returns of strategies based on the ensemble model's news tone signal as a function of when the trade is initiated, from one to ten days after news releases. In Figure 3, we plot the average annualized risk-adjusted VW and EW returns on the long minus short (L-S) portfolios. The portfolios are sorted among all stocks to reflect aggregate-level assimilation speed, and also sorted among subgroups of stocks with different retail ownerships or sizes to measure heterogeneous assimilation speed. At the aggregate level, as indicated by the green columns, returns exhibit a decay pattern as trading is postponed. They are the highest on Day 1, largely decrease on Day 2, and then circle around zero for the rest of days. In other words, the aggregate speed suggests that most news-based trading opportunities dissipate after two days.

This pattern varies across firm characteristics. High retail ownership stocks demonstrate both higher returns and longer persistence compared to low retail ownership stocks. The VW (EW) portfolio returns for high retail stocks stay positive until the 4^{th} (or 10^{th}) trading day, whereas for low retail stocks, the returns quickly turn negative on the 2^{nd} (or 6^{th}) trading day. Similarly, small-size firms show consistently stronger and more persistent returns than large-size firms. The positive portfolio returns extend to Day 6 for small firms, but turn negative on Day 3 (or Day 5)

for large firms.[24]

Overall, we show that the average aggregate public news assimilation speed is around 2 days, yet the speed can be prolonged when retail investors' behaviors or firm sizes create frictions in the rapid incorporation of LLM-extracted information.

5.4 LLM Signals and Trading Dynamics

Our economic intuition is that when news becomes public, sophisticated investors with advanced information processing capabilities may swiftly identify and act upon it. Drawing from theories such as Suominen (2001) and Easley and O'Hara (1987), these investors may submit large orders rapidly to use their informational advantage. This behavior, as Griffiths et al. (2000) suggest, could induce temporary price impacts, driving short-term return predictability and facilitating price discovery. Consequently, we hypothesize a positive relationship between LLM signals and the trading direction of larger trades, potentially from sophisticated investors, while smaller trades, possibly from less sophisticated investors, may exhibit a negative relationship with these signals.

Empirically, we compute the order imbalance (*Oib*) of four groups of trades, small, medium, large, and extra-large, varying in trade sizes.[25] Given each group *G*'s daily trades, their *Oib* for stock *i* on day *t* is calculated as $Oib_{i,t}^{G}=\frac{Buy_{i,t}^{G}-Sell_{i,t}^{G}}{Buy_{i,t}^{G}+Sell_{i,t}^{G}}$, measuring their trading direction. The higher (lower) the order imbalance, the more net-buying (net-selling). Then, we adopt the Fama-Macbeth

[24] In Appendix Figure A2, we further show that the assimilation speed can be slower when the news is more unexpected by market participants. We find that news followed by the highest magnitudes of stock price reaction needs at least 3 days to be fully incorporated into prices, whereas news that are less informative only needs around 1 day.

[25] Specifically, CSMAR categorizes trades as with small, medium, large, or extra-large sizes. If the size of a trade is lower than 50,000 CNY, then it is a small-size trade. If the size 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 is higher than or equal to 200,000 CNY but lower than one million CNY, then it is a large trade. If the size is greater than or equal to 1 million CNY, then it is an extra-large trade.

(1973) method to predict next-day order imbalance using previous-day news tones or return forecasts from the ensemble model. The equation is similar to Equation (5), where we replace the dependent variables with the next-day order imbalances for the four types of trade. In addition to the controls in Equation (5), we additionally control for the previous day's order imbalance (*Loib*) to allow for trading persistence, as in Boehmer et al. (2021) and Jones et al. (2023).

In Table 8 Panel A, for the extra-large trades, news tone has a significantly positive coefficient of 6.49 with a *t*-Stat of 8.24. For large trades, news tone also has a significantly positive coefficient of 4.43 with a *t*-Stat of 5.39. However, for small trades, news tone's coefficient is significantly negative of -4.62 with a *t*-Stat of -8.26. Similarly, in Panel B, the higher the return forecasts, the more that large traders net-buy on the next day, where the smallest traders net-sell and act as counterparties. Overall, these results imply LLM signals capture information that is more readily incorporated into trading decisions by larger, possibly more sophisticated, market participants.

6. Robustness and Further Discussion

6.1 Transaction Costs

While our primary focus is to illustrate the economic implications of LLM signals rather than optimize trading strategies, we acknowledge the importance of practical implementation. We thus extend our analysis to incorporate realistic trading costs. This includes a 10.0 bps stamp fee for selling and a 1.5 bps commission fee for both buying and selling, typical of Chinese markets. Table 9 Panel A shows the after-fee returns for trading strategies constructed among different size subsamples of stocks, and with multiple holding horizons from one to ten days.

First, we observe an annualized return decrease of nearly 60% due to the transaction costs

associated with daily rebalancing under the 1-day holding period, compared to the before-fee results documented in Section 5.2. For example, when VW, large-cap stocks' long minus short after-fee returns are -0.10%, while their before-fee returns in Table 7 are 56.17%. However, when considering VW small-cap stocks, or EW portfolios in both large-cap and small-cap stocks, although the after-fee long minus short returns are 60% lower than before-fee, they still maintain significantly positive even under daily rebalancing. For instance, VW small-cap stocks' after-fee long minus short returns are 49.13% with a *t*-Stat of 4.12 (before-fee being 108.52%), and EW after-fee long minus short returns are 15.57% (*t*-Stat of 1.96) and 52.11% (*t*-Stat of 4.29) respectively for large- and small-cap stocks. Second, recall that Section 5.4 documents that certain investors, particularly those engaging in large and extra-large trades, trade in line with LLM signals in the next day. The positive after-fee long minus short returns under 1-day holding period indicate that these investors may profit from such trading. Third, the after-fee results show that small-cap stocks continue to outperform the large-cap. Specifically, for the 1-day holding period, small-cap long minus short portfolios generate significant returns of 49.13% and 52.11% for VW and EW respectively, while large-cap portfolios yield lower returns of -0.10% and 15.57%. This outperformance persists across longer horizons, with small-caps maintaining economically and significantly long minus short returns of 22.73% (VW) and 27.78% (EW) even at 10-day holding periods. Fourth, as the holding period extends, turnover substantially decreases from around 90% to under 10%. This brings a gradual increase in the long-leg-only portfolio's after-fee returns.

6.2 LLMs and ChatGPT

Our study primarily focuses on open-sourced LLMs. However, the widespread adoption of

ChatGPT, a closed-sourced LLM, inspires the following comparative analysis. For open-sourced LLMs, they offer transparency in model structure and algorithmic details, which facilitates full result replicability, and enables further fine-tuning to incorporate specific Chinese contexts. For closed-sourced LLMs, e.g., ChatGPT, they have the advantage of rapidly equipping with advanced architectures and expanded training materials. However, such proprietary models lack transparency in training information, and face challenges in result reproducibility.[26] Given these advantages and disadvantages, we aim to compare how a leading proprietary model, ChatGPT, performs against the open-sourced in interpreting Chinese financial news and predicting returns.

Empirically, we form news tone and return forecast signals from ChatGPT by conducting similar economic modelling process using the article-level representations, which we directly retrieve from the API endpoint based on ChatGPT3.5 architecture.[27] The results are in Table 9 Panel B. First, when using news tone, no matter value-weighted or equal-weighted, ChatGPT demonstrates positive predictive power, but substantially falls short of our open-sourced ensemble model. Take the CH4-adjusted returns for instances. ChatGPT's long-minus-short portfolios achieve annualized returns of 42.07% (*t*-Stat of 4.30) when value-weighted, and 62.72% (*t*-Stat of 8.68) when equal-weighted. These returns are much lower than those of our ensemble model, which are 66.75% (value-weighted) and 91.32% (equal-weighted), respectively. The differences of 24.68% (=66.75%-42.07%) and 28.60% (=91.32%-62.72%) are economically significant.

[26] Specifically, researchers pose a concern on ChatGPT's potential look-ahead bias. The supervised fine-tuning process in ChatGPT's training may uses future stock price information. As the process is not publicly disclosed, this potential cannot be definitively ruled out. Second, the generative nature of ChatGPT's responses leads to reproducibility issues. Its outputs are not entirely deterministic, varying between identical queries or slightly altered input phrasing. This output's uncertainty, combined with frequent overwritten updates of model versions, both impede result reproducibility.
[27] Specifically, we use the "text-embedding-3-small" model from the API.

Second, when using the return forecast signal, ChatGPT continues to underperform our ensemble model in equal-weighted portfolios. It is only in value-weighted portfolios sorted by return forecast where ChatGPT slightly outperforms the ensemble by around 4% annualized. Overall, while ChatGPT shows certain promise, our ensemble of open-sourced LLMs generally maintain an edge.

Given the superior performance of open-sourced models, we further employ word cloud visualization to gain more insights into the open-sourced model's advantage in interpreting news.[28] Figure 4 reveals that frequently occurring terms in highest news tone portfolios (e.g., "激励"(incentives), "同比增长" (year-on-year growth)) align with positive sentiment indicators in prior literature. Meanwhile, terms in lowest news tone portfolios (e.g., "亏损"(losses), "减持"(selling holdings)) well correspond to negative sentiment indicators. This data inspection provides further evidence that our mainly focused LLMs can be helpful in understanding the news.

6.3 LLMs: Foreign VS. Domestic

We extend our analysis by differentiating between foreign-originated LLMs adapted to Chinese (BERT, FinBERT, RoBERTa), and domestically refined Chinese LLMs (Baichuan, ChatGLM, InternLM). We hypothesize that the latter may better capture subtle cultural references, idiomatic expressions, and market-specific jargon in Chinese financial news. Empirically, we create separate ensemble models for each group of LLMs. In Table 9 Panel C, the domestic ensemble consistently outperforms the foreign ensemble in both value-weighted and equal-weighted portfolios sorted by news tone and return forecast signals. For instance, in news tone-

[28] Specifically, we pool news from the highest or lowest news tone portfolios. Next, we conduct tokenization (using part-of-speech tagging from *jieba*), and report the frequencies of tokens. We visualize the most frequent tokens in a word cloud, with the font size illustrating the number of times they appear.

based portfolios, the domestic ensemble achieves higher CH4-adjusted annualized alphas (57.01% value-weighted, 85.41% equal-weighted) compared to the foreign ensemble (48.50% value-weighted, 76.08% equal-weighted), with similar patterns observed in return forecast-based portfolios. These findings suggest that while foreign-originated models exhibit strong general language understanding, they may fall short of domestic LLMs in specific cultural and linguistic insights, which can be crucial for interpreting Chinese financial news.

7. 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 and return forecasts from LLMs, with the ensemble model and Baichuan model generating the best market-beating cumulative returns. The significant L-S portfolio alphas hint market inefficiency for arbitrage opportunities. Second, we illustrate LLM signals can be helpful in the price discovery process of the public news. The LLM signals contain fundamental information under-processed from the news. LLMs can also be more helpful given higher frictions for market participants efficiently processing news. These frictions may originate from firms' opaque information environment and arbitrage costs, from the holdings by retail investors, and from news with complex content and central media source. The assimilation speed of the LLM-captured information is about two days at aggregate level, while varying depending on firm characteristics. Regarding how the LLM-captured information gets incorporated into prices, we suggest that heterogeneous investors trading could be a channel. Finally, we provide a battery of robustness checks and model comparisons.

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 texts, 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.

Our study hopes to serve 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 AI 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*.

Cai, Zheng, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin, 2024, InternLM2 Technical Report, *arXiv preprint arXiv: 2403.17297*.

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.

Chan, Kalok, and Allaudeen Hameed, 2006, Stock Price Synchronicity and Analyst Coverage in Emerging Markets, *Journal of Financial Economics* 80, 115–147.

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, 2019, 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.*

Easley, David, and Maureen O'Hara, 1987, Price, Trade Size, and Information in Securities Markets, *Journal of Financial Economics* 19, 69–90.

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, Yumin Liu and Lingchao Meng, 2024, Large Language Models, Textual Sentiment and Financial Markets, *Management World* 40, 42-64. *(in Chinese)*

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.

Liao, Li, Zhengwei Wang, Jia Xiang, Hongjun Yan, and Jun Yang, 2021, User Interface and Firsthand Experience in Retail Investing, *The Review of Financial Studies* 34, 4486-523.

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.

Qin, Bei, David Strömberg, and Yanhui Wu, 2018, Media Bias in China, *American Economic Review* 108, 2442–2476.

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.

Suominen, Matti, 2001, Trading Volume and Information Revelation in Stock Markets, *The Journal of Financial and Quantitative Analysis* 36, 545.

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.

Wang Kemin, Huajie Wang, Dongdong Li, and Xingyun Dai, 2018, Complexity of Annual Reports and Management Self-interest: Empirical Evidence from Chinese Listed Firms, *Management World* 34, 120-132. (*in Chinese*)

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*.

You, Jiaxing, Bohui Zhang, and Le Zhang, 2018, Who Captures the Power of the Pen?, *The Review of Financial Studies* 31, 43–96.

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*.

Zhou, Yang, Jianqing Fan and Lirong Xue, 2024, How Much Can Machines Learn Finance from Chinese Text Data?, *Management Science, forthcoming*.

Table 1. Summary Statistics

In this table, we report the summary statistics of Chinese news articles and stock characteristics. 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 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. Panel B presents the distribution of stock returns, market capitalization, EP ratio, and daily turnover.

Panel A. Number of Chinese News Articles

	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. Distribution of Stock Characteristics

Variables	Mean	Std	P25	P50	P75
Return	0.06%	3.85%	-1.39%	0.00%	1.36%
Size (billion CNY)	11.78	52.40	1.89	3.74	8.08
EP Ratio	0.48%	3.18%	0.08%	0.45%	0.97%
Turnover	2.85%	3.43%	0.91%	1.76%	3.46%

Table 2. Model Fits

In this table, we report the in-sample and out-of-sample statistical performances of the BERT, FinBERT, RoBERTa, Baichuan, ChatGLM, InternLM, and their ensemble model. In Panel A, we present the distribution of the news tone and return forecast signals during training sample. In Panel B, we present the distribution of these signals during testing sample. In Panel C, we compute the classification accuracy of the news tones provided by each model in the testing sample. In Panel D, we calculate the time-series average of the cross-sectional rank-correlations between the return forecasts and the next-day open-to-open returns for each model in the testing sample.

Panel A. In-sample Distribution of News Tone and Return Forecast Signals

		BERT	FinBERT	RoBERTa	Baichuan	ChatGLM	InternLM	Ensemble
News tone	Mean	-0.0182	-0.0182	-0.0182	-0.0182	-0.0182	-0.0181	-0.0182
News tone	Std	0.0108	0.0212	0.0205	0.0272	0.0272	0.0257	0.0193
Return forecast	Mean	-0.0001	-0.0001	-0.0001	-0.0001	-0.0001	-0.0001	-0.0001
Return forecast	Std	0.0019	0.0026	0.0025	0.0037	0.0038	0.0034	0.0025

Panel B. Out-of-sample Distribution of News Tone and Return Forecast Signals

		BERT	FinBERT	RoBERTa	Baichuan	ChatGLM	InternLM	Ensemble
News tone	Mean	-0.0772	-0.0682	-0.0727	-0.0577	-0.0604	-0.0595	-0.0660
News tone	Std	0.0443	0.0513	0.0489	0.0774	0.1043	0.0945	0.0565
Return forecast	Mean	0.0002	0.0002	0.0001	0.0000	-0.0001	0.0000	0.0001
Return forecast	Std	0.0016	0.0017	0.0017	0.0029	0.0042	0.0038	0.0024

Panel C. Out-of-Sample Classification Accuracy Using News Tones

BERT	FinBERT	RoBERTa	Baichuan	ChatGLM	InternLM	Ensemble
52.71%	52.77%	52.63%	52.37%	51.93%	52.08%	52.74%

Panel D. Out-of-Sample Cross-Sectional Correlations Using Return Forecasts

BERT	FinBERT	RoBERTa	Baichuan	ChatGLM	InternLM	Ensemble
1.62%	1.94%	1.80%	1.99%	1.52%	1.73%	1.95%

Table 3. Performance of Daily News Tone Portfolios

The table reports the performance of value-weighted (VW) and equal-weighted (EW) long-minus-short portfolios sorted by news tones and their long and short legs. The decile portfolios are built from various LLMs, including BERT, FinBERT, RoBERTa, Baichuan, ChatGLM, InternLM, and the individual 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	VW						EW					
	Long Leg		Short Leg		Long minus Short		Long Leg		Short Leg		Long minus Short	
	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat
BERT	15.90%	1.59	-31.55%	-2.73	47.45%	4.60	26.51%	2.63	-38.75%	-3.21	65.26%	9.13
FinBERT	14.07%	1.28	-37.02%	-3.00	51.08%	4.52	34.28%	3.23	-43.28%	-3.30	77.55%	8.99
RoBERTa	21.30%	2.02	-33.58%	-2.83	54.88%	4.93	30.39%	2.94	-43.87%	-3.53	74.26%	9.52
Baichuan	27.74%	2.39	-38.79%	-3.35	66.54%	5.57	41.11%	3.80	-43.28%	-3.47	84.40%	9.27
ChatGLM	18.58%	1.67	-18.79%	-1.77	37.37%	3.42	31.21%	2.99	-33.07%	-2.84	64.28%	7.98
InternLM	16.55%	1.53	-18.54%	-1.68	35.09%	3.24	38.74%	3.80	-33.97%	-2.78	72.71%	8.73
Ensemble	20.16%	1.77	-43.48%	-3.77	63.64%	5.30	39.39%	3.66	-49.12%	-3.94	88.52%	9.88

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

Model	VW						EW					
	Long Leg		Short Leg		Long minus Short		Long Leg		Short Leg		Long minus Short	
	Alpha	*t*-Stat	Alpha	*t*-Stat	Alpha	*t*-Stat	Alpha	*t*-Stat	Alpha	*t*-Stat	Alpha	*t*-Stat
BERT	9.18%	1.54	-38.22%	-5.28	47.40%	4.97	16.33%	3.74	-49.79%	-9.41	66.13%	9.83
FinBERT	7.67%	1.20	-43.77%	-5.61	51.44%	4.92	24.68%	5.09	-54.26%	-8.47	78.95%	9.68
RoBERTa	14.56%	2.23	-39.89%	-5.33	54.45%	5.22	20.32%	4.61	-54.89%	-9.42	75.22%	10.25
Baichuan	22.92%	3.66	-46.99%	-5.92	69.90%	6.52	32.42%	6.75	-54.68%	-8.64	87.09%	10.91
ChatGLM	13.60%	2.23	-27.90%	-4.13	41.50%	4.23	22.53%	4.96	-44.70%	-8.31	67.23%	9.41
InternLM	11.68%	1.85	-26.46%	-3.70	38.14%	3.83	30.02%	6.93	-45.38%	-7.57	75.41%	10.09
Ensemble	15.08%	2.41	-51.67%	-6.80	66.75%	6.42	30.78%	6.49	-60.54%	-10.03	91.32%	11.76

Table 4. Performance of Portfolios Sorted by Return Forecasts

The table reports the performance of value-weighted (VW) and equal-weighted (EW) long-minus-short portfolios and their long and short legs. The portfolios are built based on BERT, FinBERT, RoBERTa, Baichuan, ChatGLM, InternLM, and the ensemble model, respectively, using LLMs' return forecasts as the 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	VW						EW					
	Long Leg		Short Leg		Long minus Short		Long Leg		Short Leg		Long minus Short	
	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat
BERT	19.39%	1.69	-16.13%	-1.48	35.52%	3.50	27.11%	2.45	-29.38%	-2.47	56.49%	7.46
FinBERT	20.00%	1.73	-25.64%	-2.29	45.64%	4.45	35.05%	3.24	-38.16%	-3.22	73.21%	9.40
RoBERTa	20.47%	1.76	-17.54%	-1.54	38.01%	3.68	28.14%	2.49	-38.21%	-3.23	66.35%	8.51
Baichuan	14.91%	1.33	-26.21%	-2.39	41.12%	4.06	37.49%	3.42	-38.68%	-3.13	76.17%	9.36
ChatGLM	15.75%	1.32	-19.76%	-1.81	35.51%	3.46	33.70%	3.05	-33.26%	-2.77	66.96%	7.93
InternLM	17.32%	1.61	-16.33%	-1.40	33.65%	3.18	35.70%	3.25	-33.25%	-2.66	68.95%	8.33
Ensemble	15.45%	1.34	-32.08%	-2.94	47.52%	4.48	37.07%	3.37	-43.13%	-3.51	80.20%	9.24

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

Model	VW						EW					
	Long Leg		Short Leg		Long minus Short		Long Leg		Short Leg		Long minus Short	
	Alpha	*t*-Stat	Alpha	*t*-Stat	Alpha	*t*-Stat	Alpha	*t*-Stat	Alpha	*t*-Stat	Alpha	*t*-Stat
BERT	13.79%	2.03	-24.39%	-3.65	38.18%	3.85	18.08%	3.56	-41.18%	-8.13	59.25%	8.19
FinBERT	14.97%	2.23	-34.12%	-5.02	49.10%	4.94	26.31%	5.20	-50.36%	-9.99	76.67%	10.71
RoBERTa	14.70%	2.27	-25.34%	-3.70	40.04%	4.00	18.98%	3.54	-50.13%	-10.25	69.11%	9.36
Baichuan	10.59%	1.69	-34.98%	-5.44	45.57%	5.04	29.16%	5.31	-50.74%	-9.55	79.90%	10.80
ChatGLM	10.54%	1.53	-27.96%	-4.30	38.51%	3.97	25.08%	4.44	-44.93%	-8.48	70.01%	8.86
InternLM	11.57%	1.91	-24.30%	-3.19	35.88%	3.60	26.76%	4.91	-45.02%	-8.05	71.78%	9.40
Ensemble	10.80%	1.67	-40.95%	-5.92	51.75%	5.28	28.85%	5.17	-55.35%	-10.31	84.20%	10.55

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 each LLM. In Panel B, the key independent variable, *RetForecast*, is the return forecast estimated by each LLM. LLMs include BERT, FinBERT, RoBERTa, Baichuan, ChatGLM, InternLM, 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		InternLM		Ensemble	
	Coef	*t*-Stat	*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	0.0076	8.99	0.0166	10.86
Lret	-0.0147	-4.14	-0.0145	-4.04	-0.0147	-4.09	-0.0154	-4.29	-0.0154	-4.32	-0.0154	-4.27	-0.0151	-4.21
Lwret	-0.0051	-3.05	-0.0050	-2.98	-0.0051	-3.04	-0.0051	-3.01	-0.0052	-3.09	-0.0052	-3.06	-0.0051	-3.04
Lmret	-0.0005	-0.61	-0.0005	-0.63	-0.0005	-0.56	-0.0006	-0.70	-0.0007	-0.74	-0.0006	-0.70	-0.0006	-0.74
Lsize	0.0000	-0.79	0.0000	-0.81	0.0000	-0.96	0.0000	-1.01	0.0000	-0.73	0.0000	-0.76	0.0000	-1.13
Lep	-0.0012	-0.33	-0.0021	-0.59	-0.0023	-0.63	-0.0026	-0.71	-0.0010	-0.26	-0.0006	-0.18	-0.0033	-0.92
Lturn	-0.0124	-3.25	-0.0123	-3.25	-0.0127	-3.30	-0.0128	-3.41	-0.0125	-3.35	-0.0127	-3.41	-0.0124	-3.32
Intercept	0.0019	5.08	0.0016	4.35	0.0017	4.66	0.0011	3.19	0.0008	2.31	0.0009	2.43	0.0016	4.41
Adj.R2	0.04%		0.09%		0.07%		0.12%		0.09%		0.11%		0.12%	

Panel B. Returns Forecasts Predicting Returns

	BERT		FinBERT		RoBERTa		Baichuan		ChatGLM		InternLM		Ensemble	
	Coef	*t*-Stat	*Coef*	*t*-Stat	*Coef*	*t*-Stat	*Coef*	*t*-Stat	*Coef*	*t*-Stat	*Coef*	*t*-Stat	*Coef*	*t*-Stat
RetForecast	0.4660	10.20	0.4844	10.73	0.4798	11.69	0.3218	11.22	0.1798	10.27	0.2175	10.23	0.4341	11.67
Lret	-0.0150	-4.24	-0.0149	-4.24	-0.0149	-4.23	-0.0153	-4.34	-0.0153	-4.33	-0.0152	-4.29	-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	-3.25	-0.0055	-3.28
Lmret	-0.0005	-0.61	-0.0006	-0.70	-0.0006	-0.64	-0.0007	-0.76	-0.0006	-0.71	-0.0007	-0.75	-0.0007	-0.76
Lsize	0.0000	-0.58	0.0000	-0.76	0.0000	-0.69	0.0000	-0.94	0.0000	-0.80	0.0000	-0.81	0.0000	-0.98
Lep	-0.0001	-0.02	-0.0009	-0.25	-0.0006	-0.18	-0.0016	-0.46	-0.0002	-0.06	-0.0007	-0.19	-0.0017	-0.47

Lturn	-0.0129	-3.35	-0.0123	-3.19	-0.0128	-3.37	-0.0125	-3.29	-0.0121	-3.17	-0.0125	-3.25	-0.0123	-3.22
Intercept	0.0003	0.78	0.0002	0.68	0.0003	0.89	0.0004	1.10	0.0004	1.18	0.0004	1.11	0.0004	0.99
Adj.R2	0.06%		0.07%		0.07%		0.08%		0.08%		0.09%		0.09%	

Table 6. Earnings Surprise Prediction

This table reports the OLS regression results for predicting future earnings surprises using previous news tones and return forecasts. 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. In Panel A, *Tone* is the news tone provided by the ensemble model of various LLMs. In Panel B, *RetForecast* is the return forecast provided by the ensemble model. The timing of news tones and return forecasts is on 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 double clustered at stock and calendar quarter level.

Panel A. News Tones Predicting Future SUE

Dep.Var	Next-day SUE	
	Coef	*t*-Stat
Tone	4.63	8.65
Lret	1.74	4.96
Lwret	1.29	6.35
Lmret	0.68	5.37
Lsize	0.02	4.59
Lep	0.51	1.87
Lturn	-0.7	-1.42
Intercept	0.49	8.00
Adj.R2	5.38%	

Panel B. Return Forecasts Predicting Future SUE

Dep.Var	Next-day SUE	
	Coef	*t*-Stat
RetForecast	135.31	6.95
Lret	1.61	4.68
Lwret	1.17	6.99
Lmret	0.68	5.16
Lsize	0.02	4.66
Lep	0.44	1.91
Lturn	-0.72	-1.48
Intercept	0.23	4.66
Adj.R2	6.81%	

Table 7. Return Prediction for Heterogeneous Firms and News

This table reports the heterogeneity results for different firms and news. In Panel A, we consider heterogeneity for stocks with different retail holdings and information frictions. We first sort stocks into two subgroups based on each of the following characteristics. For retail ownership, "Low Retail Ownership" (or "High Retail Ownership") denotes the subgroup of stocks with lower-than-median (or higher-than-median) percentage of shares held by retail investors in the previous quarter. For market capitalization, "Large-Cap" (or "Small-Cap") denotes the higher-than-median (or lower-than-median) size subgroup. For 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. For state-ownership, "State Owned" (or "Non State Owned") is an indicator for whether the stock is state-owned or not. For analyst coverage, "Nonzero Analyst Coverage" (or "Zero Analyst Coverage") is an indicator for whether the stock has nonzero analyst coverage in the previous calendar year that issue earnings forecasts. Within each subgroup, we then sort by daily news tones or return forecasts based on the ensemble model of various LLMs, and form value-weighted or equal-weighted decile portfolios. Decile portfolios with the lowest (or highest) news tones or return forecasts are denoted by the Short Leg (or Long Leg). Long minus short portfolio denotes the trading strategy that buys the long-leg and shorts the short-leg. In Panel B and Panel C, we consider heterogeneity at the news level. In Panel B, we first sort stocks into two subgroups based on their corresponding news' adoption of uncommon Chinese characters. "W/ Uncommon Characters" (or "W/O/ Uncommon Characters") denotes the subgroup of news that use (or do not use) uncommon Chinese characters. The dictionary of common characters is defined in "The List of Common Modern Chinese Characters" (a total of 3,500 characters) published by the State Language Work Committee in 1988. Within each subgroup, we then sort by daily news tones or return forecasts based on the ensemble model of various LLMs, and form value-weighted or equal-weighted decile portfolios. Columns "Ret" and "*t*-Stat" stand for each portfolio's annualized return and *t*-Statistics. In Panel C, we consider whether the news is reported by central media. We follow the definition for central media from Cyberspace Administration of China's "List of Internet News Information Sources". The dependent variable is the next-day stock return. *CentralMedia* is a dummy variable that takes the value of 1 when the news comes from a central media. Control variables are similar to Table 5. We include both stock- and day-level fixed effects.

Panel A. Heterogeneous Firms

Subgroup	VW						EW					
	Long Leg		Short Leg		Long minus Short		Long Leg		Short Leg		Long minus Short	
	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat
Sorted by News Tones												
Low Retail Ownership	24.64%	2.04	-37.08%	-3.02	61.72%	4.53	40.25%	3.68	-40.97%	-3.28	81.22%	7.55
High Retail Ownership	20.82%	1.63	-55.21%	-3.87	76.02%	5.35	39.20%	3.23	-55.14%	-3.91	94.34%	8.61
Large-cap	24.20%	2.02	-31.97%	-2.78	56.17%	4.67	33.35%	3.11	-38.50%	-3.23	71.85%	8.02
Small-cap	45.57%	3.42	-62.95%	-4.11	108.52%	8.73	54.05%	4.03	-57.52%	-3.67	111.57%	8.89
Nonzero Shorting	23.66%	1.96	-22.76%	-1.97	46.42%	3.85	37.05%	3.24	-39.46%	-3.36	76.51%	7.71
Zero Shorting	38.36%	2.84	-66.43%	-4.56	104.79%	7.96	49.94%	4.11	-57.54%	-4.08	107.47%	10.12
State Owned	11.02%	0.83	-25.87%	-1.77	36.89%	2.47	22.58%	2.01	-36.17%	-2.57	58.76%	5.08
Non State Owned	28.71%	2.36	-46.63%	-3.52	75.33%	5.81	47.14%	4.19	-47.78%	-3.59	94.92%	8.93
Nonzero Analyst Coverage	23.19%	2.00	-23.42%	-2.03	46.60%	4.00	38.77%	3.53	-36.21%	-3.07	74.98%	9.01
Zero Analyst Coverage	12.40%	0.89	-75.99%	-4.28	88.38%	5.30	38.82%	2.79	-65.79%	-3.75	104.61%	6.56
Sorted by Return Forecasts												
Low Retail Ownership	11.30%	0.93	-30.13%	-2.64	41.43%	3.27	40.24%	3.54	-33.36%	-2.82	73.59%	7.63
High Retail Ownership	18.90%	1.33	-28.30%	-1.92	47.20%	3.39	37.14%	2.88	-50.88%	-3.56	88.02%	8.04
Large-cap	13.71%	1.18	-24.55%	-2.27	38.25%	3.67	20.28%	1.83	-30.40%	-2.68	50.68%	5.74
Small-cap	47.73%	3.41	-54.51%	-3.68	102.24%	8.06	58.92%	4.17	-51.69%	-3.52	110.61%	8.83
Nonzero Shorting	19.52%	1.61	-13.64%	-1.17	33.16%	2.81	20.17%	1.73	-22.91%	-1.90	43.08%	4.11
Zero Shorting	22.87%	1.64	-60.71%	-4.26	83.57%	6.45	49.49%	3.91	-56.54%	-4.21	106.03%	10.24
State Owned	3.44%	0.28	-21.54%	-1.64	24.98%	1.75	15.59%	1.43	-32.49%	-2.38	48.08%	4.44
Non State Owned	23.30%	1.79	-31.70%	-2.57	55.00%	4.31	46.73%	3.79	-48.24%	-3.83	94.97%	9.06
Nonzero Analyst Coverage	16.01%	1.37	-21.15%	-1.98	37.17%	3.62	34.40%	3.06	-26.38%	-2.34	60.78%	7.24
Zero Analyst Coverage	22.34%	1.45	-75.23%	-4.36	97.58%	5.74	49.47%	3.24	-68.57%	-3.98	118.04%	7.30

Panel B. Heterogeneous News Complexity

Subgroup	VW						EW					
	Long Leg		Short Leg		Long minus Short		Long Leg		Short Leg		Long minus Short	
	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat
Sorted by News Tones												
W/ Uncommon Characters	43.33%	3.39	-54.96%	-4.18	98.30%	6.53	52.30%	4.53	-57.48%	-4.38	109.79%	9.51
W/O/ Uncommon Characters	11.73%	0.97	-37.00%	-2.81	48.73%	3.62	30.11%	2.66	-45.79%	-3.22	75.90%	6.67
Sorted by Return Forecasts												
W/ Uncommon Characters	35.40%	2.76	-46.69%	-3.59	82.09%	6.24	56.65%	4.52	-53.25%	-4.17	109.90%	9.21
W/O/ Uncommon Characters	-5.43%	-0.40	-22.39%	-1.73	16.95%	1.15	21.87%	1.87	-38.18%	-2.96	60.05%	6.00

Panel C. Heterogeneous News Source

Dep.Var	Next-day return		Next-day return	
Signal	News Tone		Return Forecast	
	Coef	*t*-Stat	*Coef*	*t*-Stat
Signal	0.0142	11.94	0.3075	12.62
Signal*CentralMedia	0.0106	2.79	0.3019	2.83
CentralMedia	0.0005	1.41	-0.0004	-1.83
Lret	-0.0218	-5.08	-0.0224	-11.13
Lwret	-0.0058	-2.97	-0.006	-6.38
Lmret	-0.001	-1.06	-0.001	-2.51
Lsize	-0.0002	-8.64	-0.0002	-3.14
Lep	-0.0022	-1.61	-0.0021	-1.38
Lturn	-0.0226	-5.72	-0.0228	-11.90
Intercept	0.0023	15.25	0.0013	9.00
Adj.R2	0.18%		0.17%	

Table 8. Trading Order Imbalances Prediction

This table reports the Fama-Macbeth regression results for predicting next-day trading order imbalances using previous day news tones and return forecasts. The dependent variables are the next-day order imbalances for four types of trades: trades with small 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 larger the trade sizes, the more aggressive the trades are. Trade size data are obtained from CSMAR database. Order imbalance (*Oib*) for a specific type of trade is calculated as that type's number of buy trades minus sell trades over the sum of the number of buy and sell trades. In Panel A, *Tone* is the news tone provided by the ensemble model of various LLMs. In Panel B, *RetForecast* is the return forecast provided by the ensemble model. 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.

Panel A. News Tones Predicting Heterogeneous Order Imbalances

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	-4.62	-8.26	-0.43	-0.82	4.43	5.39	6.49	8.24
Loib	0.23	71.16	0.13	41.36	0.06	21.46	0.01	4.49
Lret	-8.55	-8.67	4.32	4.33	8.60	5.45	-1.58	-0.97
Lwret	0.90	1.71	6.05	11.44	8.08	10.94	-2.04	-2.84
Lmret	0.77	2.41	3.67	12.83	4.69	10.32	0.26	0.63
Lsize	0.03	4.52	0.16	16.40	0.16	14.36	-0.02	-2.53
Lep	5.92	3.09	10.40	5.22	-1.69	-0.70	-13.64	-4.90
Lturn	3.62	2.85	2.19	1.80	-5.79	-2.67	-9.13	-5.15
Intercept	-0.26	-1.94	-4.20	-30.20	-5.01	-21.54	0.64	4.54
Adj.R2	5.12%		1.80%		0.24%		0.01%	

Panel B. Return Forecasts Predicting Heterogeneous Order Imbalances

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
RetForecast	-67.68	-5.08	112.31	8.67	190.82	9.61	145.93	7.62
Loib	0.23	71.18	0.13	41.28	0.06	21.34	0.01	4.43
Lret	-8.03	-8.18	4.64	4.92	8.54	5.51	-1.59	-0.96
Lwret	1.01	1.90	6.02	11.31	7.83	10.38	-2.01	-2.82
Lmret	0.83	2.56	3.68	13.04	4.64	10.43	0.29	0.70
Lsize	0.03	4.39	0.16	16.24	0.16	14.50	-0.02	-2.54
Lep	5.28	2.77	8.88	4.58	-2.40	-1.00	-12.60	-4.76
Lturn	3.62	2.84	2.24	1.85	-5.51	-2.50	-9.23	-5.25
Intercept	0.06	0.50	-4.17	-31.74	-5.34	-24.02	0.17	1.40
Adj.R2	5.08%		1.83%		0.23%		0.00%	

Table 9. Performances Considering Transaction Costs, ChatGPT, and Recombined Ensemble Models

The table reports portfolio performances when transactions costs are accounted for, and when using ChatGPT or recombined ensemble models. In Panel A, we consider the after-fee performances of long minus short portfolios and their long and short legs, depending on different holding horizons and subsamples of stocks. We deduct a stamp fee of 10.0 bps (upon selling) and commission fee of 1.5 bps (upon both buying and selling) for every transaction. Then we report the after-fee CH4-adjusted annualized returns, value-weighted (VW) or equal-weighted (EW), for portfolios sorted by news tones from the ensemble model. The portfolios are held for multiple horizons including 1-day, 5-day, and 10-day, and they are constructed among stocks with higher- or lower-than-median market-caps. In Panel B, we report the performance of value-weighted (VW) and equal-weighted (EW) long minus short portfolios and their long and short legs, sorted by signals from ChatGPT model. In Panel C, we report the performance of value-weighted (VW) and equal-weighted (EW) long minus short portfolios and their long and short legs for two recombined ensemble models. We separate the six individual LLMs into two groups: foreign-originated and linguistically adapted to Chinese (including BERT, FinBERT, and RoBERTa) and domestic-modified models (including Baichuan, ChatGLM, and InternLM). We create an ensemble model for each group by equal-weighting constituent models' signals. The portfolios are built based on the two ensemble models' news tones or return forecasts. "Ret" and "*t*-Stat" stand for each portfolio's annualized raw or CH4-adjusted returns and *t*-Statistics.

Panel A. Portfolio Performances After Accounting for Transaction Costs

			Long Leg			Short Leg			Long minus Short		
Weighting	Group	Holding Days	Ret	*t*-Stat	Turnover	Ret	*t*-Stat	Turnover	Ret	*t*-Stat	Turnover
VW	Large cap	1-day	-9.69%	-1.40	89.78%	-9.59%	-1.25	90.50%	-0.10%	-0.01	90.14%
VW	Large cap	5-day	5.35%	0.68	18.48%	-3.33%	-0.40	18.83%	8.68%	1.22	18.66%
VW	Large cap	10-day	7.27%	1.07	9.31%	-0.03%	0.00	9.50%	7.30%	1.27	9.41%
VW	Small cap	1-day	3.83%	0.49	95.26%	-45.30%	-4.82	89.32%	49.13%	4.12	92.29%
VW	Small cap	5-day	16.58%	1.92	19.35%	-18.99%	-1.94	18.75%	35.57%	4.92	19.05%
VW	Small cap	10-day	17.60%	2.36	9.70%	-5.13%	-0.59	9.49%	22.73%	4.03	9.59%
EW	Large cap	1-day	-4.00%	-0.79	92.28%	-19.56%	-3.06	88.97%	15.57%	1.96	90.62%
EW	Large cap	5-day	7.58%	1.01	18.94%	-9.24%	-1.10	18.63%	16.82%	3.03	18.79%
EW	Large cap	10-day	8.52%	1.26	9.52%	-2.98%	-0.39	9.43%	11.50%	2.49	9.48%
EW	Small cap	1-day	12.06%	1.50	95.15%	-40.05%	-4.16	88.83%	52.11%	4.29	91.99%
EW	Small cap	5-day	22.47%	2.53	19.34%	-19.71%	-1.95	18.66%	42.17%	5.26	19.00%
EW	Small cap	10-day	21.62%	2.82	9.70%	-6.15%	-0.70	9.45%	27.78%	4.52	9.57%

Panel B. Portfolio Performances Based on Signals from ChatGPT Model

	VW						EW					
	Long Leg		Short Leg		Long minus Short		Long Leg		Short Leg		Long minus Short	
	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat
News tone	15.51%	1.43	-25.66%	-2.23	41.17%	3.88	28.77%	2.87	-32.66%	-2.65	61.43%	7.97
News tone, CH4-adjusted	9.07%	1.45	-33.00%	-4.75	42.07%	4.30	19.19%	4.48	-43.53%	-7.53	62.72%	8.68
Return forecast	24.25%	2.24	-28.06%	-2.52	52.31%	5.47	35.00%	3.29	-30.72%	-2.56	65.72%	8.60
Return forecast, CH4-adjusted	18.27%	3.11	-36.64%	-5.29	54.91%	5.93	25.93%	5.11	-43.11%	-8.56	69.04%	9.60

Panel C. Recombined Ensemble Models, CH4 Adjusted Alpha

		VW						EW					
		Long leg		Short leg		Long minus Short		Long leg		Short leg		Long minus Short	
		Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat
News tone	Foreign	7.75%	1.21	-40.75%	-5.67	48.50%	4.84	22.45%	4.8	-53.63%	-8.97	76.08%	10.05
News tone	Domestic	17.33%	2.73	-39.68%	-5.25	57.01%	5.4	31.23%	6.79	-54.18%	-9.07	85.41%	11.03
Return forecast	Foreign	14.61%	2.12	-28.13%	-3.94	42.74%	4.05	24.50%	4.58	-48.28%	-9.32	72.79%	9.61
Return forecast	Domestic	9.64%	1.56	-43.47%	-6.31	53.11%	5.41	29.80%	5.31	-55.84%	-10.56	85.64%	10.98

Figure 1. An Example of Deriving the Article-level Representation

This figure provides an example of deriving the article-level representation using the Baichuan model. Given a piece of example news article, the model first tokenizes the article. For each token, the embedding matrix of the model transforms the token into a high-dimensional vector. The article-level representation, $X_{i,t}$, is derived by taking the average of the embedding vectors.

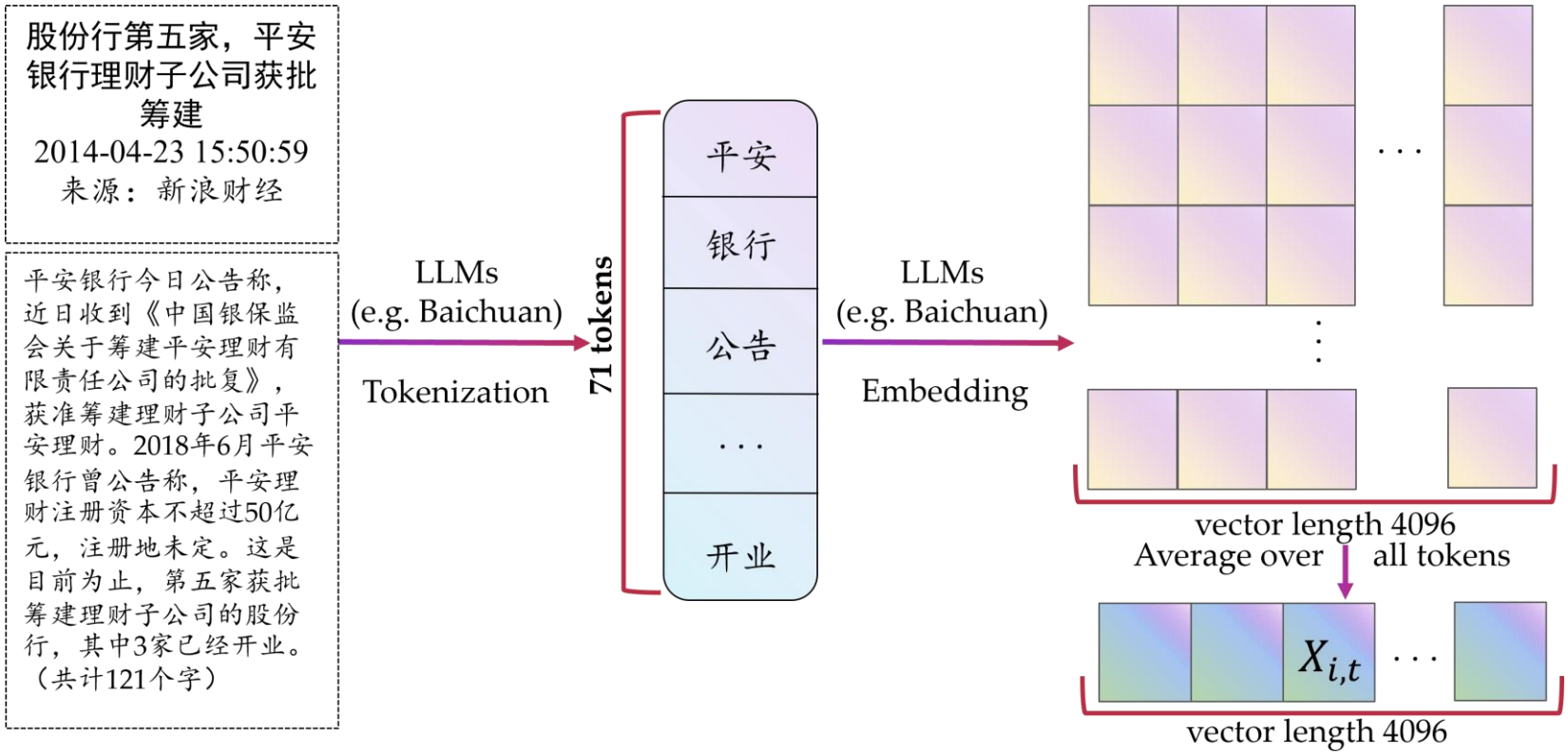
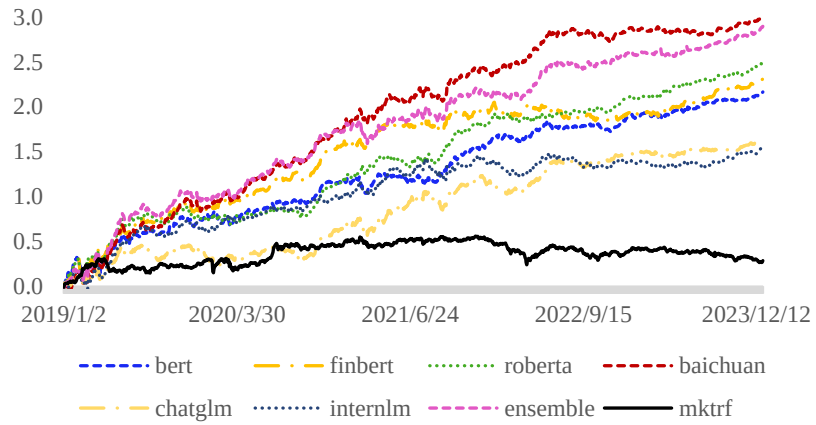


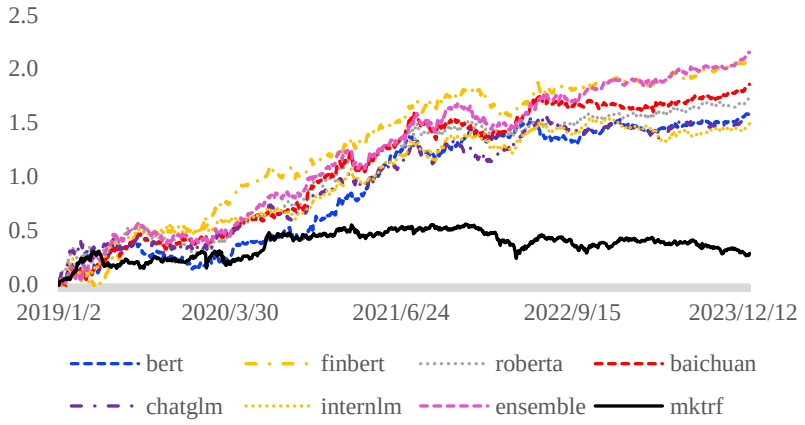
Figure 2. Portfolio Performances of Various LLMs

This figure plots the value-weighted and equal-weighted cumulative log returns for long-minus-short portfolios, sorted by news tones or return forecasts from BERT, FinBERT, RoBERTa, Baichuan, ChatGLM, InternLM, and the ensemble model. "Mkt" represents the cumulative A-share market return.

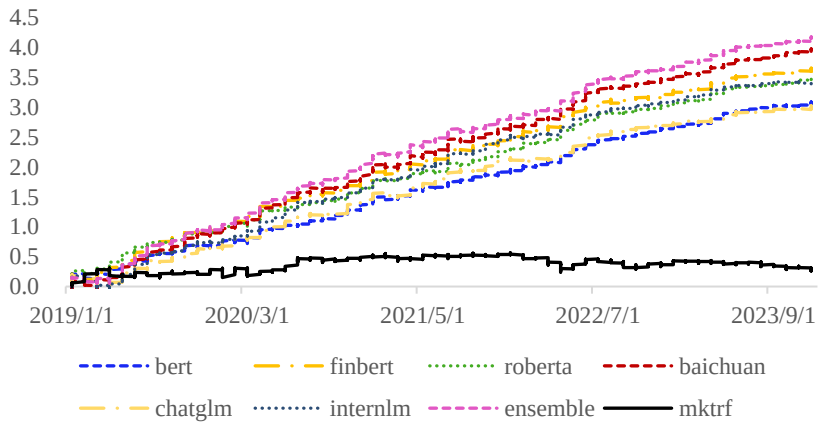
Panel A. Value-Weighted Portfolios Sorted by News Tones



Panel B. Value-Weighted Portfolios Sorted by Return Forecasts



Panel C. Equal-Weighted Portfolios Sorted by News Tones



Panel D. Equal-Weighted Portfolios Sorted by Return Forecasts

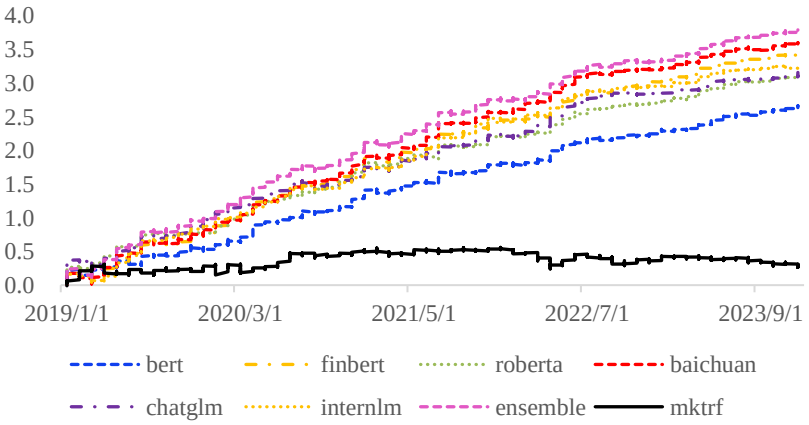
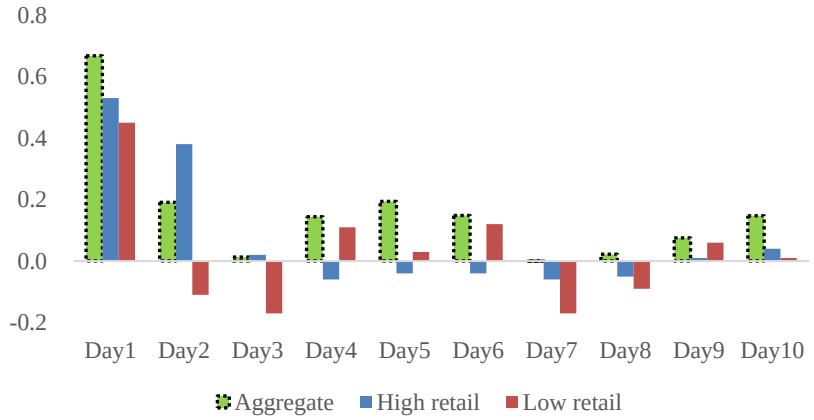


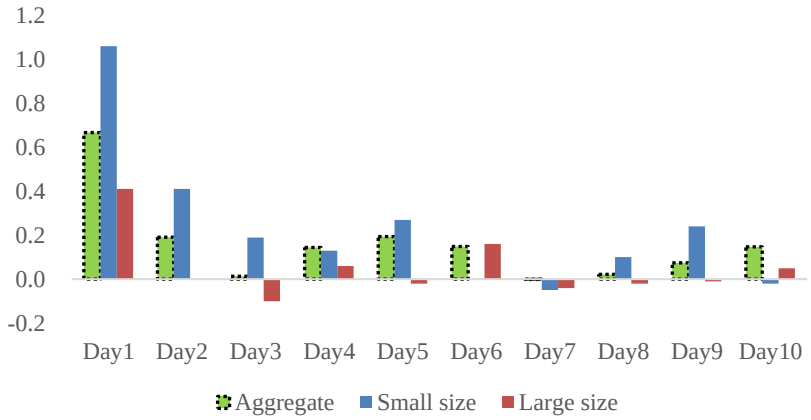
Figure 3. Speed of News Assimilation

This figure compares average one-day returns to the trading strategy based on the ensemble model, as a function of when the trade is initiated, from one to ten days following the news. Panel A (or Panel C) reports value-weighted (or equal-weighted) average annualized CH4-adjusted returns on the long minus short portfolios, sorted by news tones among all stocks, or subgroups of stocks with higher- or lower-than-median retail ownerships. Panel B (or Panel D) reports value-weighted (or equal-weighted) results sorted by news tones among all stocks, or subgroups of stocks with smaller- or larger-than-median sizes.

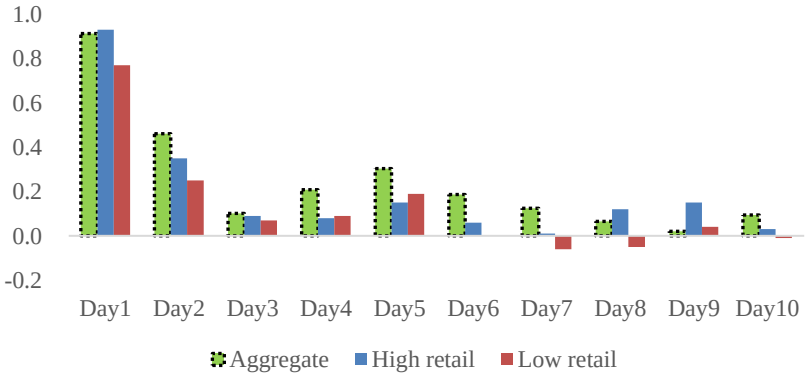
Panel A. Value-Weighted Portfolios for Retail Subgroups



Panel B. Value-Weighted Portfolios for Size Subgroups



Panel C. Equal-Weighted Portfolios for Retail Subgroups



Panel D. Equal-Weighted Portfolios for Size Subgroups

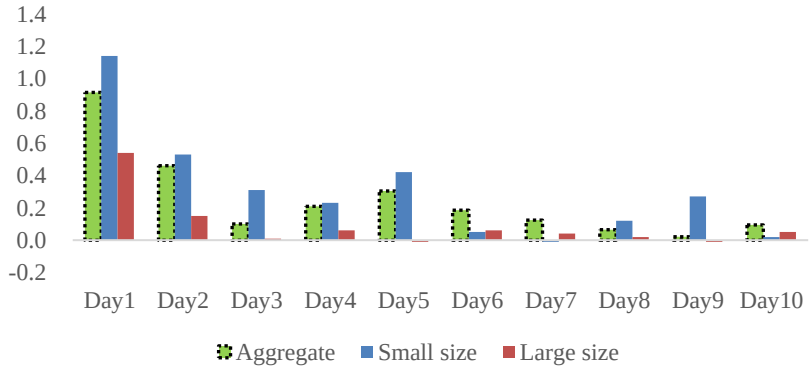


Figure 4. Chinese Word Clouds of Long and Short Portfolios
This figure plots the word clouds of long and short portfolios using the ensemble model.

Panel A. Long Portfolios' Word Cloud

Panel B. Short Portfolios' Word Cloud

Appendix 1. Details on Large Language Models

LLMs 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 econometric modelling in Section 3.1. 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 word 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 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, e.g., BERT, 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 unlabeled data, and (2) it enables transfer learning by initializing models with 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. It 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. Additional Tables and Figures

Appendix Table A1. Number of Characters/Tokens in Chinese News Articles

In this table, we 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", "# of ChatGLM Tokens", and "# of InternLM Tokens" report the percentiles of the number of tokens converted from news text using model specific tokenizer.

	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
# of InternLM Tokens	42	97	152	226	530

Appendix Table A2. Sharpe Ratios of Daily News Tone Portfolios in China

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

Panel A. Sharpe Ratios of Value-weighted Portfolios

Model	Long-leg only			Short-leg only			Long minus Short		
	Ret	Std	SR	Ret	Std	SR	Ret	Std	SR
BERT	15.90%	22.89%	0.69	-31.55%	24.77%	-1.27	47.45%	21.81%	2.18
FinBERT	14.07%	24.14%	0.58	-37.02%	26.28%	-1.41	51.08%	24.24%	2.11
RoBERTa	21.30%	23.33%	0.91	-33.58%	26.33%	-1.28	54.88%	23.54%	2.33
Baichuan	27.74%	26.11%	1.06	-38.79%	25.16%	-1.54	66.54%	26.12%	2.55
ChatGLM	18.58%	26.00%	0.71	-18.79%	23.86%	-0.79	37.37%	23.71%	1.58
InternLM	16.55%	25.86%	0.64	-18.54%	23.49%	-0.79	35.09%	23.76%	1.48
Ensemble	20.16%	26.20%	0.77	-43.48%	25.13%	-1.73	63.64%	25.94%	2.45

Panel B. Sharpe Ratios of Equal-weighted Portfolios

Model	Long-leg only			Short-leg only			Long minus Short		
	Ret	Std	SR	Ret	Std	SR	Ret	Std	SR
BERT	26.51%	22.86%	1.16	-38.75%	24.12%	-1.61	65.26%	14.82%	4.40
FinBERT	34.28%	23.59%	1.45	-43.28%	24.95%	-1.73	77.55%	17.37%	4.46
RoBERTa	30.39%	23.21%	1.31	-43.87%	24.39%	-1.80	74.26%	15.71%	4.73
Baichuan	41.11%	23.99%	1.71	-43.28%	24.37%	-1.78	84.40%	18.25%	4.62
ChatGLM	31.21%	23.37%	1.34	-33.07%	23.75%	-1.39	64.28%	16.19%	3.97
InternLM	38.74%	23.29%	1.66	-33.97%	23.82%	-1.43	72.71%	16.27%	4.47
Ensemble	39.39%	23.98%	1.64	-49.12%	24.68%	-1.99	88.52%	17.80%	4.97

Appendix Table A3. Performance Robustness of Daily News Tone Portfolios

The table reports the robust performance of value-weighted (VW) and equal-weighted (EW) portfolios sorted by news tones, after filtering out stocks suspended from trading or hitting the price limit when forming the portfolios. The decile portfolios are sorted by news tones from BERT, FinBERT, RoBERTa, Baichuan, ChatGLM, InternLM, and ensemble model. "Ret", "Alpha" and "*t*-Stat" stand for each portfolio's annualized raw and CH4-adjusted return and *t*-Statistics.

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

Model	VW						EW					
	Long-leg only		Short-leg only		Long minus Short		Long-leg only		Short-leg only		Long minus Short	
	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat
BERT	15.22%	1.51	-27.30%	-2.37	42.52%	4.19	23.40%	2.34	-33.58%	-2.75	56.98%	8.23
FinBERT	14.73%	1.34	-32.81%	-2.66	47.54%	4.24	31.22%	2.96	-40.75%	-3.13	71.97%	8.73
RoBERTa	20.63%	1.95	-28.86%	-2.46	49.48%	4.52	27.50%	2.66	-43.74%	-3.59	71.24%	9.37
Baichuan	27.31%	2.35	-32.93%	-2.83	60.23%	4.96	31.82%	3.01	-41.39%	-3.39	73.21%	8.65
ChatGLM	17.87%	1.59	-16.60%	-1.54	34.47%	3.23	22.63%	2.17	-32.10%	-2.75	54.73%	7.34
InternLM	16.01%	1.48	-17.27%	-1.56	33.28%	3.11	29.20%	2.91	-34.01%	-2.88	63.21%	8.34
Ensemble	19.69%	1.72	-37.19%	-3.16	56.88%	4.76	29.99%	2.84	-46.67%	-3.79	76.67%	9.17

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

Model	VW						EW					
	Long-leg only		Short-leg only		Long minus Short		Long-leg only		Short-leg only		Long minus Short	
	Alpha	*t*-Stat	Alpha	*t*-Stat	Alpha	*t*-Stat	Alpha	*t*-Stat	Alpha	*t*-Stat	Alpha	*t*-Stat
BERT	8.43%	1.41	-34.46%	-5.12	42.88%	4.59	13.12%	3.17	-45.11%	-9.01	58.23%	8.98
FinBERT	8.28%	1.29	-39.82%	-5.30	48.09%	4.60	21.51%	4.67	-52.20%	-8.77	73.71%	9.47
RoBERTa	13.81%	2.12	-35.74%	-5.06	49.55%	4.80	17.32%	4.01	-55.17%	-10.37	72.49%	10.12
Baichuan	22.41%	3.62	-41.86%	-5.38	64.27%	5.93	22.95%	5.47	-53.32%	-9.24	76.27%	10.24
ChatGLM	12.83%	2.09	-25.93%	-3.92	38.76%	4.08	13.77%	3.26	-44.14%	-8.91	57.91%	8.75
InternLM	10.92%	1.73	-25.61%	-3.66	36.53%	3.70	20.27%	5.33	-45.72%	-8.54	66.00%	9.72
Ensemble	14.55%	2.33	-45.97%	-6.09	60.52%	5.84	21.17%	4.99	-58.55%	-10.49	79.73%	10.98

Appendix Table A4. Performance Robustness of Portfolios Sorted by Return Forecasts

The table reports the robust performance of value-weighted (VW) and equal-weighted (EW) portfolios sorted by return forecasts, after filtering out stocks suspended from trading or hitting the price limit when forming the portfolios. The decile portfolios are sorted by the return forecast signals from BERT, FinBERT, RoBERTa, Baichuan, ChatGLM, InternLM, and ensemble model. "Ret", "Alpha" and "*t*-Stat" stand for each portfolio's annualized raw and CH4-adjusted return and *t*-Statistics.

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

Model	VW						EW					
	Long-leg only		Short-leg only		Long minus Short		Long-leg only		Short-leg only		Long minus Short	
	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat
BERT	16.66%	1.46	-13.82%	-1.27	30.48%	3.01	13.68%	1.28	-29.61%	-2.50	43.29%	6.06
FinBERT	17.40%	1.50	-22.32%	-2.00	39.72%	3.93	22.19%	2.14	-37.47%	-3.16	59.66%	8.08
RoBERTa	19.16%	1.66	-17.30%	-1.51	36.46%	3.51	15.50%	1.42	-36.17%	-3.06	51.67%	7.10
Baichuan	10.48%	0.93	-23.92%	-2.16	34.40%	3.46	17.56%	1.68	-36.05%	-2.95	53.61%	7.33
ChatGLM	13.02%	1.09	-16.08%	-1.52	29.10%	2.92	15.82%	1.49	-29.34%	-2.49	45.16%	6.17
InternLM	13.39%	1.24	-13.39%	-1.17	26.77%	2.54	16.95%	1.58	-29.35%	-2.37	46.29%	6.18
Ensemble	11.91%	1.04	-25.96%	-2.39	37.87%	3.65	17.33%	1.63	-39.48%	-3.20	56.81%	7.18

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

Model	VW						EW					
	Long-leg only		Short-leg only		Long minus Short		Long-leg only		Short-leg only		Long minus Short	
	Alpha	*t*-Stat	Alpha	*t*-Stat	Alpha	*t*-Stat	Alpha	*t*-Stat	Alpha	*t*-Stat	Alpha	*t*-Stat
BERT	10.93%	1.61	-22.28%	-3.37	33.21%	3.34	4.43%	0.96	-41.70%	-8.65	46.13%	6.76
FinBERT	12.26%	1.82	-30.98%	-4.54	43.23%	4.42	13.24%	3.08	-49.94%	-10.27	63.18%	9.35
RoBERTa	13.37%	2.11	-25.33%	-3.63	38.71%	3.83	6.37%	1.32	-48.32%	-10.38	54.69%	7.94
Baichuan	6.00%	0.95	-32.79%	-5.17	38.80%	4.34	8.79%	1.92	-48.21%	-9.50	56.99%	8.57
ChatGLM	7.54%	1.11	-24.28%	-3.98	31.83%	3.37	6.87%	1.50	-41.16%	-8.34	48.03%	7.04
InternLM	7.60%	1.23	-21.34%	-2.91	28.95%	2.90	7.60%	1.66	-41.30%	-7.92	48.91%	7.12
Ensemble	7.03%	1.11	-34.47%	-4.98	41.50%	4.28	8.82%	1.90	-51.74%	-9.93	60.56%	8.34

Appendix Table A5. Long-Run Return Prediction

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 signal 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	
	Coef	*t*-Stat
k=1	0.0072	6.82
k=2	0.0007	0.68
k=3	0.0003	0.32
k=4	0.0003	0.35
k=5	0.0003	0.35
k=6	0.0015	1.66
k=7	0.0011	1.33
k=8	0.0015	1.81

Appendix Table A6. Additional Results on Performance of Daily News Tone Portfolios Based on Heterogeneous News

The table reports additional results on the performance of long minus short portfolios and their long and short legs sorted by news tones, based on different news categories and characteristics. In Panel A, news categories include firm announcements (accounting for 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 decile portfolios are built based on the ensemble model. Panel B reports the performance of long minus short portfolios and their long and short legs sorted by news tones, based on different news characteristics. We consider two news characteristics: 1) negation usage; 2) number usage. "With Negation" (or "Without Negation") denotes the subgroup of news with (or without) usage of negation, while "High Negation Ratio" (or "Low Negation Ratio") denotes the subgroup of news with higher-than-median (or lower-than-median) ratio of negation usage in the previous day. "High Number Ratio" (or "Low Number Ratio") denotes the subgroup of news with higher-than-median (or lower-than-median) ratio of numbers in the previous day. The decile portfolios are built based on the ensemble model of various LLMs. "Ret" and "*t*-Stat" stand for each portfolio's annualized return and *t*-Statistics.

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

News Category	VW						EW					
	Long-leg only		Short-leg only		Long minus Short		Long-leg only		Short-leg only		Long minus Short	
	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat
Firm Announcements	14.12%	1.03	-64.58%	-4.56	78.70%	5.10	47.31%	3.81	-50.63%	-3.55	97.95%	7.71
Operation News	6.25%	0.47	-17.08%	-1.26	23.33%	1.49	20.90%	1.82	-20.00%	-1.61	40.89%	3.68
Equity News	36.18%	2.34	-46.80%	-3.06	82.98%	4.34	52.97%	4.18	-31.44%	-2.21	84.41%	5.88

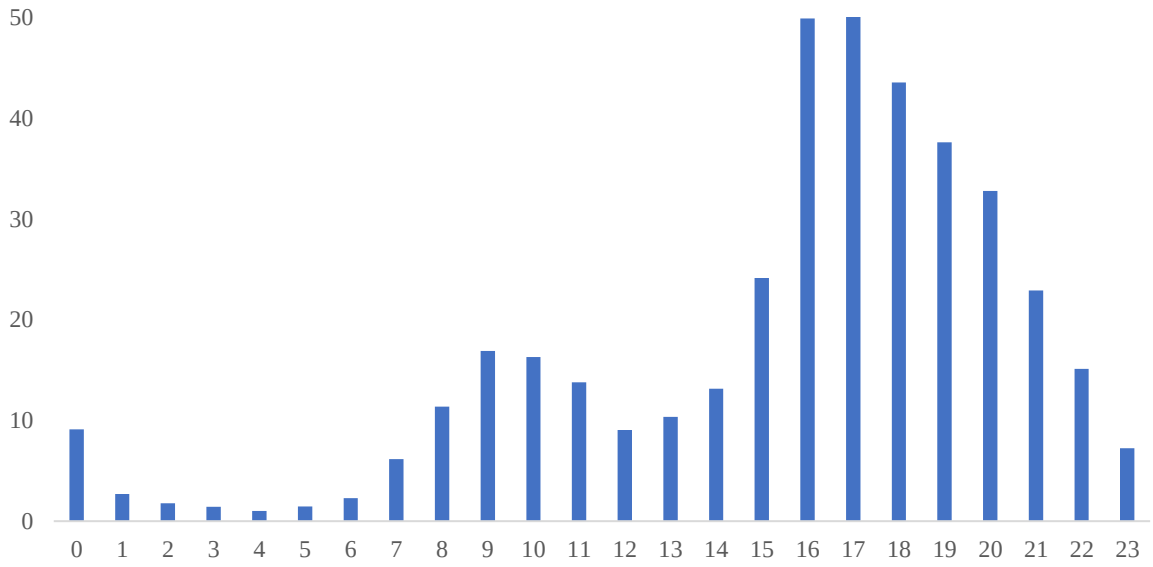
Panel B. News Differing in Adoption of Negation and Number

Subgroup	VW						EW					
	Long Leg		Short Leg		Long minus Short		Long Leg		Short Leg		Long minus Short	
	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat	Ret	*t*-Stat
Sorted by News Tones												
With Negation	24.97%	2.12	-62.65%	-4.4	87.62%	6.33	39.27%	3.33	-61.30%	-4.35	100.56%	9.25
Without Negation	29.20%	2.18	-19.24%	-1.5	48.45%	3.2	48.28%	3.96	-39.42%	-3.16	87.70%	7.16
High Negation Ratio	25.83%	2.25	-58.35%	-3.95	84.19%	5.89	35.62%	3.07	-58.81%	-4.11	94.43%	8.44
Low Negation Ratio	28.15%	2.17	-20.31%	-1.62	48.46%	3.39	49.25%	4.15	-36.72%	-2.96	85.98%	7.5

High Number Ratio	26.06%	2.2	-44.67%	-3.28	70.73%	4.81	42.53%	3.9	-44.05%	-3.31	86.58%	7.49
Low Number Ratio	27.71%	2.1	-41.22%	-2.9	68.92%	4.54	41.20%	3.32	-48.28%	-3.5	89.48%	7.91
Sorted by Return Forecasts												
With Negation	24.61%	1.87	-53.05%	-4.12	77.66%	5.90	37.32%	3.13	-62.89%	-4.81	100.21%	9.55
Without Negation	15.76%	1.22	-3.67%	-0.29	19.43%	1.40	44.97%	3.59	-28.89%	-2.25	73.87%	6.35
High Negation Ratio	22.74%	1.69	-47.24%	-3.63	69.98%	5.10	34.02%	2.80	-57.73%	-4.38	91.75%	8.27
Low Negation Ratio	13.58%	1.08	-8.52%	-0.69	22.10%	1.67	44.95%	3.54	-29.97%	-2.41	74.92%	6.89
High Number Ratio	25.40%	1.99	-32.87%	-2.51	58.28%	3.95	34.82%	3.09	-47.81%	-3.62	82.63%	7.40
Low Number Ratio	11.10%	0.83	-39.10%	-2.92	50.20%	3.86	46.43%	3.39	-46.77%	-3.36	93.20%	7.44

Appendix Figure A1. Chinese News Count: Intraday Pattern

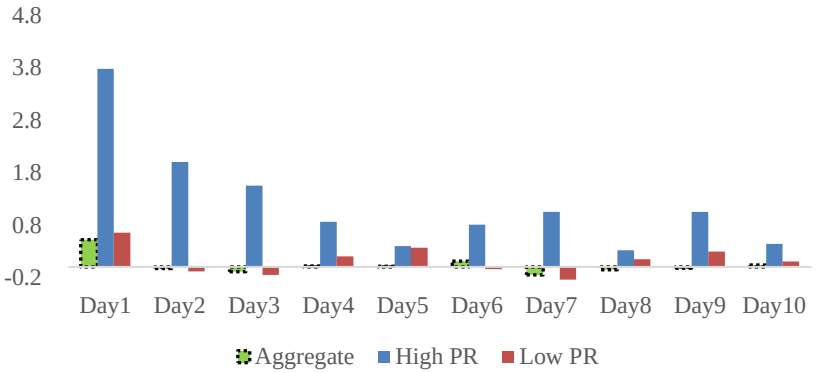
This figure plots the average number of news articles per hour (24-hour local time) in China, from January 2008 to December 2023.



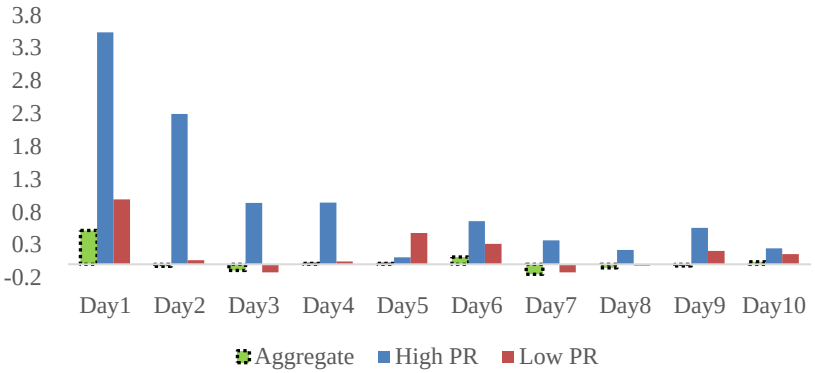
Appendix Figure A2. Heterogeneous Speed of News Assimilation

This figure presents heterogeneous news assimilation speed depending on stock price reaction (denoted as PR hereafter). We employ a double-sorting method. In Panel A and C (B and D), we first sort by the ensemble model's news tones (return forecast) signals, and then double-sort by stocks' return reactions following the news, to form value-weighted and equal-weighted portfolios. High (low) reaction sub-decile is denoted as "High PR" ("Low PR"). We plot average one-day holding period annualized CH4-adjusted returns to the long minus short portfolios, as a function of when the trade is initiated, ranging from one to ten days following the news.

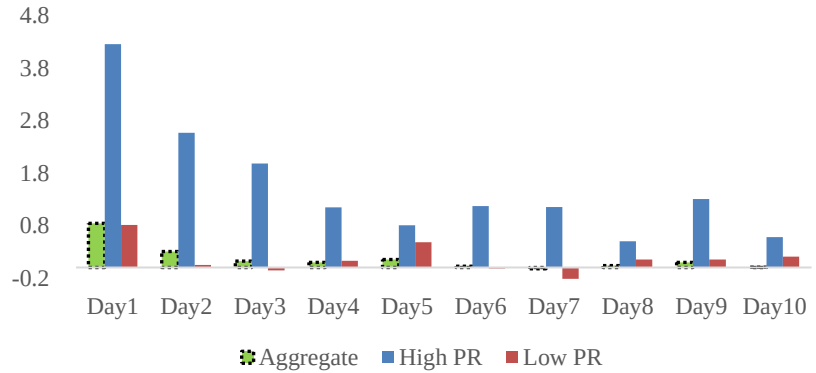
Panel A. Value-Weighted Portfolios Sorted by News Tones



Panel B. Value-Weighted Portfolios Sorted by Return Forecasts



Panel C. Equal-Weighted Portfolios Sorted by News Tones



Panel D. Equal-Weighted Portfolios Sorted by Return Forecasts