Measuring Price Effects of Multilingual Global News with Large Language Models

Kevin Benson* Ing-Haw Cheng†‡ Maurice Granger* John Hull†
Charles Martineau† Yoshio Nozawa† Vasily Strela* Yuntao Wu† Jun Yuan*†

February 25, 2026

Abstract

We study how language boundaries affect price discovery by using multilingual language models to decompose Bloomberg news into language-specific ('unaligned') sentences and content-matched ('aligned') sentences. For Japanese equities, stock prices underreact to unaligned English-only sentences consistent with domestic investors underweighting information available only in foreign-language coverage. Stock prices also underreact to English and Japanese sentences whose content is aligned. In U.S. equities, stock prices underreact only to aligned English coverage. Overall, the evidence suggests that markets do not efficiently aggregate reporting across languages, creating predictable returns for investors who process multilingual information streams.

Keywords: Large language models, international finance, textual analysis

*Royal Bank of Canada (RBC) Capital Markets. New York, NY and Toronto, ON, Canada.

†Rotman School of Management, FinHub Financial Innovation Lab, University of Toronto. 105 St. George Street, Toronto, ON, M5S 3E6, Canada.

‡Corresponding author. The authors thank seminar participants at RBC Quantitative Trading Strategies, Hong Kong University, Hong Kong University of Science and Technology, and City University of Hong Kong for comments and feedback. The authors also thank RBC Capital Markets for generous funding of the FinHub: Financial Innovation Lab and Haruka Takagi, Sean Woo, Ray Fang, Zixuan Lu, Weixuan Jiang, and Lynn Tao for providing excellent research assistance. We also thank Bloomberg News for providing us news data and Junichi Yamanoi for the Japanese foreign ownership data. Ing-Haw Cheng thanks the Social Science and Humanities Research Council of Canada (SSHRC) for generous funding through SSHRC Insight Grant #435-2024-0365. The views expressed in this article are those of the authors and do not necessarily reflect those of the Royal Bank of Canada.

How well do markets around the world price news that appears in multiple languages? Global news services such as Bloomberg provide round-the-clock coverage of international markets in various languages. In principle, this should give global investors access to the same information regardless of language. In practice, investors read the language of their choice, nuances make translation imperfect, and editorial decisions can cause coverage to differ subtly across languages. If these language boundaries matter, then prices may incorporate news scattered across languages at different speeds, and traders could profit from paying attention to news in multiple languages. To what extent modern markets are effectively "omnilingual" and how language frictions shape price discovery, are ultimately empirical questions.

This paper studies how language barriers shape price discovery in Japanese and U.S. equities. We build on prior work by distinguishing two potential channels. First, some investors may discount or miss content in news reported only in one language. Second, investors may not efficiently aggregate news across languages even when content is similar across languages. In both cases, prices may not react efficiently due to language barriers.

To help separate these channels, we apply multilingual large language models (LLMs) to fifteen years of Bloomberg articles covering Japanese and U.S. stocks. We separate individual sentences across Japanese and English articles into those with content that appears only in one language ("unaligned" sentences) and those that report similar content in both languages ("aligned"), and predict returns out-of-sample. If unaligned sentences predict returns, language matters because investors discount content available only in one language. If aligned sentences also predict, language boundaries may introduce frictions even when the content is similar.

We find evidence of both channels, indicating that markets are far from "omnilingual": investors not only miss news reported in only one language, but also do not efficiently use the same news reported across multiple languages to sharpen their views. Cross-sectional tests using data on stock characteristics and article readership further support the existence of language barriers.

Section 1 introduces our data and methodology. We assemble a novel panel of Bloomberg news for Japanese and U.S. equities from 2008–2024, covering both Japanese- and English-language articles and their readership on the terminal. Using multilingual, language-agnostic

sentence embeddings, we represent every sentence of the news as a vector in a common semantic space. We then distinguish content shared across languages from language–specific content by applying an optimal-transport matching algorithm to identify pairs of sentences with similar embeddings. Sentences that match across languages form two 'aligned' groups (one per language) for that stock-day. Sentences with no credible counterpart form two 'unaligned' groups that capture language-specific content.

We construct return scores by training supervised models to predict next-day open-to-close returns from embeddings calculated from all text, only the aligned sentences, and only the unaligned sentences. The resulting return scores are generated in an out-of-sample, rolling fashion to avoid look-ahead bias and serve as language-specific trading signals based on the Bloomberg news flow.

Section 2 documents a baseline level of return predictability from Japanese and English news scores prior to applying our content separation method. In Japan, long–short portfolios formed on Japanese and English news scores both deliver large and statistically significant alphas, with an edge for the local language and robustness across several embedding models. Cross-sectionally, both Japanese and English scores remain powerful predictors of next-day returns even after controlling for size and beta. For U.S. stocks, both English and Japanese scores predict returns. While these findings imply substantial underreaction to news, consistent with Chen, Kelly, and Xiu (2022), they do not resolve whether different language feeds predict returns because they cover different events.

Section 3 applies our methodology to decompose this predictability into aligned and unaligned channels. We first develop a conceptual framework for "language-segmented processing" to guide interpretation. In the model, Japanese and English investors treat non-native-language news as noisier or, more generally, misperceive the precision of signals across languages. This motivates two empirical questions. First, do language-specific unaligned sentences forecast returns, indicating that prices do not fully incorporate content from one language stream? Second, do aligned sentences with similar content forecast returns, indicating that prices do not fully aggregate the aligned information across languages?

Empirically, we find evidence supporting affirmative answers to these questions. For Japanese stocks, English-only unaligned sentences robustly predict subsequent returns, con-

sistent with domestic investors underweighting information that appears only in foreign-language coverage. Aligned Japanese and English sentences, which are highly similar in content, also predict returns, indicating that markets do not fully aggregate overlapping news across languages.

Heterogeneity tests around size and foreign ownership support a language-based interpretation. An alternative explanation for the predictability of both English and Japanese aligned sentences is that the econometrician observes two noisy measurements of the same event even if many investors efficiently aggregate information across languages. While we cannot fully rule out this hypothesis, we find that the predictability of aligned sentences is strongest where language barriers are likely to bind most tightly: smaller, more domestically owned firms. Moreover, the pricing of aligned news is more inefficient when readership of articles within that same language is low.

Results for Japanese-only unaligned sentences show a more mixed pattern. In the full sample, slopes are typically close to zero in baseline specifications, but become negative in the subsample of large, high–foreign-ownership firms or when Japanese-language readership is high. In contrast, in the post-2020 robustness analysis the UJ point estimates are more consistently positive. While the negative slopes are consistent with generalized misperception of signal precision across languages (e.g., overconfidence in Japanese-only details), we caution against interpreting any one slope in isolation.

Taken together, the evidence points to language boundaries as a systematic friction in price discovery: markets are not "omnilingual." Investors in Japan tend to overlook English-specific content and do not fully integrate content-matched news across languages. In contrast, in U.S. stocks, only aligned English scores predict returns, suggesting that the aligned English news captures a significant piece of the predictive power for Japanese and English news for stock returns.

Section 4 reports a series of robustness checks and discusses practical implications. We show that our main results survive restricting the sample to post-2020 news. This analysis alleviates the serious concerns raised in recent work about look-ahead bias when using LLMs in finance (e.g., Levy, 2024) since the training period for the LLM we use ends in 2020. Our results also survive using stricter sentence-alignment rules and replacing our baseline

embedding model with alternative multilingual encoders. From a practical perspective, multilingual large language models can be used to build return-forecasting signals directly from raw text, without hand-crafted dictionaries; foreign-language coverage of a stock is valuable even when it appears to report the "same" story as local news; and sentence-level alignment can be used to flag genuinely new, language-specific information. Section 5 concludes.

Literature review. Our main contribution is to provide empirical evidence that language boundaries slow price discovery even when underlying news content is highly similar. While extensive literature documents stock price underreaction to news, such as earnings announcements in earlier historical periods (Ball and Brown, 1968; Bernard and Thomas, 1989; Martineau, 2022) and slow information diffusion (Hong and Stein, 1999), the literature has also established that language boundaries affect investor behavior and contribute to inefficiency. Grinblatt and Keloharju (2001) show that Finnish investors prefer stocks that communicate in their local language and are geographically close, contributing to home bias. Du et al. (2022) develop a Chinese financial sentiment dictionary and show that it outperforms translated English dictionaries, underscoring the importance of language in sentiment analysis. Lang et al. (2025) find that firms disclose information in different languages to cater to local investors, suggesting that they recognize the existence of language frictions. Durand, Khuu, and Smales (2023) analyzes Japanese and English-language Thomson Reuters sentiment indices and finds that the two measures diverge significantly, but both predict market returns, especially before 2013. Our analysis builds on this literature by separating "what is said" from "what language it is in" and shows that the latter matters: even highly correlated return scores derived from semantically aligned sentences in different languages predict returns independently.

We also contribute to the literature on applications of artificial intelligence (AI), machine learning (ML), and textual analysis to financial markets. Chen, Kelly, and Xiu (2022) shows how return scores constructed from LLM embeddings from news text can forecast stock returns across several international markets, complementing work from Gu, Kelly, and Xiu (2020) showing that AI/ML methods can extract more signal for predicting returns than classical techniques. Lopez-Lira and Tang (2025) and Chen et al. (2025) show that frontier

models such as ChatGPT can predict stock price reactions to news, with the extent of look-ahead bias an active area of research (Levy, 2024; Rahimikia and Drinkall, 2024; Sarkar and Vafa, 2024; He et al., 2025; Engelberg et al., 2025). These methods build on prior work, including applying machine learning to asset prices (Gu, Kelly, and Xiu, 2020), using text to improve forecasting (Bybee et al., 2024; Manela and Moreira, 2017), and building text-based sentiment measures (Loughran and McDonald, 2011; Tetlock, 2007). Relative to this literature, we introduce a new method to separate aligned and unaligned sentences across different languages and apply it to Bloomberg news data to understand how language-specific and shared news content predict stock returns.

1 Data and methodology

1.1 Data sources

1.1.1 News data

We collect global news articles from the Bloomberg terminal news feed covering the period starting from November 2008 to September 2024.[1] Following Bloomberg conventions, we retrieve news for the Japanese market if it covers at least one ticker with the suffix "@JP". For U.S. stocks, we retrieve news with the historical ticker that trades on AMEX, NASDAQ, or NYSE from CRSP. This procedure yields approximately 59,000 Japanese-language articles and 72,000 English-language articles per year for Japanese stocks, and 337,000 English-language articles and 12,000 Japanese-language articles per year for U.S. stocks.

To ensure we use stock-specific news, we further filter the stories using a Bloomberg News-derived ticker score. This metric indicates the relevance of a story to a specific ticker. We keep only the stories that have a single ticker with a ticker score greater than or equal to 75 out of 100, indicating median to high relevance to a specific stock traded on the market of interest. As each story may be updated multiple times, we use the last update within 24 hours of the story's publication on the Bloomberg news feed.

For each article, we remove Bloomberg-specific headers and footers and line breaks within

[1]We thank Bloomberg for providing us access to the data under a non-disclosure agreement.

a paragraph. We do not remove numbers or stopwords, since they may help preserve sentence structure when processed by machine learning models. We exclude articles shorter than 100 characters and a small fraction of articles with more than 100,000 characters. The former set may contain little information, while the latter set tends to be sprawling and diffuse in information.

1.1.2 Stock data

We retrieve Japanese stock return data from Compustat S&P Global. We compute stock returns, adjusted for splits and dividends. As in Chen, Kelly, and Xiu (2022), we use open price data to calculate a daily return from open to close. Since open prices are only available in Compustat Global starting from 2018, we retrieved them from Yahoo Finance for the prior period.[2] We retrieve U.S. stock return data from CRSP with share codes 10, 11, and 31 (American Depositary Receipts) traded on NYSE, AMEX, and NASDAQ.

1.1.3 Summary statistics

Table 1 reports the summary statistics for the total number of unique Japanese and U.S. stocks in our sample and the total number of Bloomberg articles. Panel A shows that our data cover 3,800 Japanese and 7,400 Japanese and U.S. stocks, respectively. Panel B reports that news about Japanese stocks is in both English and Japanese, with slightly fewer Japanese articles (943,000) than English articles (1.2 million). In contrast, Panel C reports that news about U.S. stocks is predominantly in English, with 5.5 million English and 194,000 Japanese articles. Upon inspection, it appears that Japanese news coverage of U.S. stocks focuses mainly on those in the S&P 500 index.

Figure 1 reports the yearly summary statistics for the number of articles, where we focus on articles with either a headline or body content for brevity. In both Japan and the U.S., the overall number of articles has declined through the years, likely related to changes in editorial processes, even as the number of stocks covered has increased. (The exception is that the number of U.S. stocks covered in Japanese language has declined in recent years.)

[2]S&P Global confirmed that the missing open price prior to 2018 is due to a change in data provider. We use the previous day's distribution and splits to adjust the open price. We confirm that the open prices from Yahoo Finance and Compustat Global match during the period of overlap.

We report in the Internet Appendix additional details on the number of words per article (for Japanese, we count the number of characters). The average length of a body for a Japanese (English) article covering a Japanese firm is 235 (160), and the average length of a Japanese (English) article covering a U.S. firm is 352 (260).

Bloomberg publishes articles 24 hours a day. Figure 2 shows the average number of articles, per stock, published in the 24 hours leading up to the open of the Japanese and U.S. markets in Panels A and B, respectively. In Japan, article production peaks 20 to 15 hours before the market opens. Production declines overnight for both English and Japanese articles, with a more pronounced drop for Japanese content. In the U.S., articles are published steadily throughout the day, with English production declining only slightly during the New York evening hours. During this same period, Japanese article production increases, coinciding with the start of the business day in Tokyo. We report in the Internet Appendix that in Japan, English articles published overnight tend to be longer in length than those published during daytime hours.

1.2 Methodology

We convert all article release times to each market's local timezone. We then align articles with stock prices and returns based on release times: articles published between 30 minutes before market open on day $t-1$ and 30 minutes before market open on day t are associated with day t's prices and returns. For example, the Japanese market opens at 9:00 am, so we link an article released at 8:28 am on day t to, in our subsequent exercise, predict returns on day t. An article released at 8:50 am on day t is associated with the stock price and return on day $t+1$. In the U.S. market, the corresponding cutoff time is 9:00 am. We retain only articles associated with trading days; Friday and weekend news are associated to the following Monday's return. Figure 3 illustrates the timeline of how we associate each day's news with returns the next day.

We use the Language-agnostic BERT sentence embedding model, or LaBSE, by Feng et al. (2022), to extract embeddings from the news text every day. Each embedding in LaBSE is a 768-dimensional vector that captures the semantic content of the input text

by analyzing the tokens in that text.[3] LaBSE is a language-agnostic encoder that assigns similar embeddings to semantically equivalent text across 109 languages. We focus primarily on LaBSE in our analysis, partly because we can confirm whether our results hold in the period after 2020, the end of its pre-training period. We also consider more recent models, such as GTE (Zhang et al., 2024) and Multilingual-E5 (Wang et al., 2024).

1.2.1 Daily JP/EN embeddings

As a benchmark, we calculate a representative English-language and a representative Japanese-language embedding for every stock day. To do so, we first input each article into LaBSE and obtain the embedding associated with it. We then average these article-level embeddings within each language to obtain separate embeddings for English and Japanese on a given stock-day. In our initial analysis, we extract information from an article, either from the headline (H), body (B), or headline+body (HB), since headlines and bodies are conceptually distinct. We thus have six sets of embeddings for each stock day: $X_{i,t}^{l,k}$ for $l \in \{EN, JP\}$ and $k \in \{H, B, HB\}$.

1.2.2 Aligning sentences using optimal transport

The process of constructing daily JP/EN embeddings may wash out subtle nuances of each sentence. To better capture how content aligns across languages, we use a finer-grained, sentence-alignment approach based on optimal transport theory, which overcomes limitations of naïve cosine similarity-based methods (which tend to be dense and uninformative) and statistical alignments (which may fail when the coverage diverges across languages).

Optimal transport (OT), initially developed by Monge (1781), provides a principled way to map the probability mass from one distribution to another while minimizing transport cost.[4] Let $\{x_i\}_{i=1}^n$ and $\{y_j\}_{j=1}^m$ denote two point sets with associated probability distributions

[3] As a BERT-based model Devlin et al. (2019), LaBSE has a maximum input length of 512 tokens. A token can be a subword, a word, or a word phrase in English, and a character or phrase in Japanese. This limit is sufficient to cover most of the articles we analyze.

[4] The original Monge (1781) formulation is often intractable and may not even admit a solution unless restrictive conditions are met, as proved by Caffarelli, Feldman, and McCann (2002). Instead, we adopt the Kantorovich (1942) formulation, and in particular its discrete version as described in Peyré and Cuturi (2020) and applied in Dou and Neubig (2021).

p_x and p_y, where $\sum_{i=1}^{n} p_{x_i} = 1$ and $\sum_{j=1}^{m} p_{y_j} = 1$. The cost of moving mass from x_i to y_j is given by $c_{ij} = c(x_i, y_j)$. The optimal transport problem seeks a transport plan $\gamma_{ij} \geq 0$ that minimizes the total cost of moving probability mass:

$$\min_{\gamma} \left\{ \sum_{ij} c_{ij}\gamma_{ij} : \gamma_{ij} \geq 0, \sum_{i} \gamma_{ij} = p_{y_j}, \sum_{j} \gamma_{ij} = p_{x_i} \right\}$$

Dou and Neubig (2021) and Swanson, Yu, and Lei (2020) show that the resulting transport plan is self-normalized and sparse, making it more effective than naïve cosine similarity comparisons. In practice, the optimal γ is approximated via entropic regularization using the Sinkhorn (1964) algorithm.[5] A large value of γ_{ij} indicates that x_i and y_j are likely to have similar semantics.

To operationalize OT within our context, we follow Dou and Neubig (2021) but extend the method from word-level to sentence-level alignments. First, for each stock i and trading day t, we concatenate all associated English articles into a single composite article $\mathcal{E}$, and all associated Japanese-language articles into $\mathcal{J}$ (we omit the stock-day subscripts to simplify notation). We use body text only and keep only the stock-days with both English and Japanese news available. We use spaCy (Honnibal et al., 2020) monolingual models to split each article into sentences, yielding $\mathcal{E} = \{E_1, ..., E_n\}$ for English articles and $\mathcal{J} = \{J_1, ..., J_m\}$ for Japanese-language articles. We then obtain embeddings for each sentence using LaBSE, normalize them, and stack them into matrices $X^E \in \mathbb{R}^{n\times 768}$ and $X^J \in \mathbb{R}^{m\times 768}$ for English and Japanese, respectively.

Next, we calculate cosine similarities for every possible English-Japanese sentence pair that will shape our cost function. Specifically, we calculate $\xi_{ij} = X_i^E \cdot X_j^J$ for every sentence i in English and j in Japanese normalized on a min-max scale to $[0, 1]$, and define our cost function as $c_{ij} = 1 - \xi_{ij}$. We then run optimal transport in both directions (English to Japanese, and Japanese to English) using these costs and starting from a uniform distribution that assigns each sentence equal probability mass, which yields two transport plans $\gamma \in \mathbb{R}^{n,m}$ and $\gamma' \in \mathbb{R}^{m,n}$. Greater values of γ_{ij} indicate a greater amount of probability mass

[5] Wang and Qiu (2025) further improved computational efficiency by exploiting sparse and low-rank matrix structures, yielding up to a tenfold speedup in our setting and making large-scale analysis feasible.

transported from English sentence i to Japanese sentence j, and vice versa for γ'_{ij}.

We use the transport plans to classify aligned sentences as follows. First, we select all entries in γ and γ' that fall within the top 10% of their overall probability mass within each transport plan, which we call *global top-mass filtering* and is a many-to-many sentence mapping. (We consider a one-to-one alignment method in Section 4.1.) Second, we require $\xi_{ij} \geq \xi_{thres}$, since different news coverage across languages may lead to spurious matches even after global top-mass filtering. The hyperparameter ξ_{thres} depends on the distribution of cosine similarities in the embedding model, and we use $\xi_{thres} = 0.6$ for LaBSE after some inspection. Sentence pairs (i, j) that satisfy both criteria are *aligned*, while those that do not are *unaligned*.

Intuitively, aligned sentences are those selected by both optimal transport plans and have a cosine similarity exceeding a minimum threshold. Figure 4 Panel A shows the alignments between English and Japanese news for Mitsubishi UFJ Financial (8306.T) on 2024-04-23. Aligned sentences have clear semantic overlap, and unaligned English sentences contain several details, such as comments from HDFC's CFO, not found in Japanese. Panel B shows alignment between English and Japanese news coverage for Apple (AAPL) on 2024-01-08. Similar to the MUFG case, the aligned sentences capture the narrative of an iPhone sales slump in China driven by the Huawei Mate 60 Pro series. The unaligned English sentences provide additional context related to TSMC and the software ecosystem.

Figure 5 plots the average proportion of sentences that our process classifies as aligned each year. In Japan, on average, 50-80% of sentences are aligned each day, with more unaligned Japanese-specific sentences in recent years. About 40% of English sentences are aligned, and the majority of sentences are unaligned. In the U.S., the pattern differs from that in Japan where only 20% of English sentences are aligned in recent years. These patterns show that there is a mix of aligned and unaligned news across different language feeds each day, and that this mix changes over time and across markets. Our process thus provides a rich source of variation for our subsequent tests.

After classifying sentences and calculating the average embedding for aligned and unaligned sentences, the final output of our process is four embeddings $(X_{i,t}^{A,E}, X_{i,t}^{A,J}, X_{i,t}^{U,E}, X_{i,t}^{U,J})$ each day representing the embeddings for Aligned/Unaligned $\times$ Japanese/English on stock-

day (i, t). Each aggregated embedding is a 768-dimensional vector.

1.2.3 Return score construction

We take our language-level embeddings (JP/EN × H/B/HB) and aligned/unaligned embeddings (A/U × JP/EN) and use them to construct daily return scores using ridge regression models following Chen, Kelly, and Xiu (2022). We target open-to-close returns given our emphasis on price discovery, denoted $R_{i,t}^{\text{Open}\to\text{Close}}$ (or just $R_{i,t}^{O\to C}$).

Specifically, we fit our ridge regression models to calculate weights w by estimating:

$$\arg\min_{w} \left\| X_{i,t}^{l,k} w - R_{i,t}^{O\to C} \right\|_2^2 + \lambda \left\| w \right\|_2^2,$$

for $\{X_{i,t}^{l,k}\}$ for $l \in \{JP, EN\}$ and $k \in \{H, B, HB, A, U\}$ and where λ is a regularization parameter selected via cross-validation over the grid 10, 20, 30,..., 100, with optimal values typically chosen as 90 or 100.

We adopt a rolling-window framework to ensure that our scores are calculated out-of-sample relative to the returns we predict in our main analysis. Specifically, for each calendar year y, the model is trained on data from years $y-5$ through y and used to generate out-of-sample return scores $\widehat{R}_{i,t}^{l} = X_{i,t}^{l,k} w$ for trading days t in year $y+1$. This procedure, repeated annually for the years 2008–2024, ensures that return scores rely exclusively on past data. The analysis period for evaluation is 2014–2024.

2 Predictability from Multilingual News Scores

In this section, we document how return scores based on Japanese- and English-language news predict short-horizon returns at the stock level. We first use portfolio sorts to provide a time–series view of performance and then move to cross-sectional regressions that control for firm characteristics. Throughout, we focus on the one–day open–to–close return, which isolates the market's reaction on the day that news is released or the following morning.

2.1 Return score summary statistics

Table 2 reports the summary statistics for our return scores, $\widehat{R}_{i,t}^{\mathrm{l,\,pre}}$, constructed from text embeddings generated in each language l using three models: LaBSE, GTE, and E5. For brevity, the table only reports summary statistics from headbody (HB) scores. Panel A shows that for Japanese stocks, the average return scores across all models and languages are negative, ranging from -0.10% to -0.13%. In contrast, the average return scores for U.S. stocks in Panel B are approximately 0%.

2.2 Time series evidence

For each trading day with at least 20 stocks with news, we sort stocks into quartiles based on the Japanese or English return scores, go long the highest–score quartile and short the lowest–score quartile, and calculate the open-to-close return. We then track the performance of each strategy.

Table 3 summarizes a simple horse race of annualized Sharpe ratios between local and foreign language news in each market. For every combination of market (Japan vs U.S.), embedding model (LaBSE, GTE, E5), and news field (headline, body, or head+body), we form daily long–short portfolios based on the corresponding return scores and compute their Sharpe ratios. Panel A reports, for each market/model cell, which language–field pair delivers the highest Sharpe ratio.

Two facts stand out. First, the local language is the top performer in five of six market-model cells, with the lone exception of when using GTE embeddings in Japan. Second, the body of the news article can offer incremental information over the headline, with three of six cells indicating that head+body outperforms headline-only. Taken together, when we look across all reasonable ways of building news-based strategies, incorporating local-language text from both the headline and body appears important in delivering the highest risk-adjusted performance. Finally, Sharpe ratios are quite large, consistent with the evidence and magnitudes from Chen, Kelly, and Xiu (2022).

Panels B–C of Table 3 report the average daily alphas of these strategies for Japanese stocks, broken out by embedding model. For each model, we report long–short alphas for

strategies formed on Japanese and English-language scores. Given the results in Panel A, we focus on head+body text.

Two patterns emerge. First, for all three embedding models, long–short portfolios based on Japanese (Panel B) and English (Panel C) scores generate sizeable and statistically significant daily alphas. Second, the Japanese (local) scores tend to perform slightly better. Within each model, the Japanese strategy delivers a similar or higher alpha than its English counterpart. The Internet Appendix reports analogous results for the U.S. market, results with transaction costs (Sharpe ratios drop by about one-half), and intraday analysis.

Figure 6 offers an alternative perspective. Panel A plots the average return to long-short strategies formed on the six language-level embedding scores (JP/EN × H/B/HB) and shows that strategies based on Japanese scores deliver greater returns than those based on English scores. Panel B plots the decay of that average return into the following day. We calculate these as the return to the same portfolio formed at the open on the portfolio formation date over successive non-overlapping trading periods. The results suggest that the scores predict underreaction to news over the first daytime session as the scores predict positive returns over the first day. There is some evidence of reversal following the first daytime session, but this is not a robust pattern across Japanese and English scores.

2.3 Cross-sectional evidence

To move beyond portfolio averages and compare language-specific signals directly in the cross-section, we estimate daily regressions of open–to–close returns on pre-market Japanese and English return scores. Table 4 reports estimates of:

$$\text{Ret}_{i,t}^{\text{Open}\to\text{Close}} = a_t + b_0 \widehat{R}_{i,t}^{\text{EN, pre}} + b_1 \widehat{R}_{i,t}^{\text{JP, pre}} + \Gamma' X_{i,t} + \varepsilon_{i,t},$$

where a_t represent date effects, $\widehat{R}_{i,t}^{\text{l,pre}}$ is the pre-market predicted return score for stock i on day t in language l, and $X_{i,t}$ includes log market capitalization measured on $t-1$ and the 3-year weekly Dimson (1979) market β. The slopes on $\widehat{R}_{i,t}^{\text{l, pre}}$ are the coefficients of interest.

For Japanese stocks, Table 4 shows that both English and Japanese scores have positive and highly economically and statistically significant slopes. Columns 1 and 2 in each panel

report the effect of each return score on returns, while Columns 3 and 4 jointly estimate the variables of interest with and without the set of control variables. In the baseline specification of column 3 without controls, the coefficient on the English score interaction is economically large and the coefficient on the Japanese score interaction is larger still. Adding size and beta in column 4 reduces both slopes by a relatively small magnitude, but they remain economically and statistically significant. This confirms the portfolio result: both languages carry incremental information about subsequent Japanese returns, with a slight advantage for the local (Japanese) scores.

For U.S. stocks, the cross-sectional patterns depend on the sample but are qualitatively similar to those in Japan. In a broad U.S. universe, English scores are the main driver of predictability, but Japanese scores continue to add predictive power even conditional on English news scores.

Overall, the language-level evidence in this section shows that there is strong underreaction to news in both the U.S. and Japan: daily strategies based on Bloomberg news in either language earn sizeable alphas, and the cross-sectional regressions show that Japanese and English scores predict subsequent returns. Local languages tend to have a modest edge, but predictability is not confined to local news: foreign-language scores also predict returns.

3 Predictability from Aligned and Unaligned News

This section decomposes the return predictability of English and Japanese news into two channels: a language-specific coverage channel, in which language-specific content predicts returns, and a content-overlap channel, in which content that we hold fixed via alignment predicts returns.

3.1 Conceptual Framework

We first develop a simple framework to help organize the subsequent interpretation of empirical evidence. We sketch the outline here and save details for the Internet Appendix.

Investors, news, and language structure. We focus on the Japanese market and assume two representative types of mean-variance investors: *Japanese investors* (J), who primarily read Japanese news, and English investors (E), who primarily read English news. Let q_J and q_E denote their relative masses with $q_J + q_E = 1$. We assume $q_J > q_E$ so that Japanese investors are the main price-setters but English investors are non-negligible.

There is a single risky asset with payoff x tomorrow which has fundamentals driven by three underlying dimensions: one associated with "dual-language" news items (e.g., an important earnings announcement), one that is more strongly reflected in Japanese-language coverage (local politics), and one that is more strongly reflected in English-language coverage (the global investor angle). The news provides four public noisy signals about these dimensions: s_{AJ} and s_{AE} are Japanese and English news about the dual-language dimension, s_{UJ} is news about the Japanese-specific dimension, and s_{UE} is news about the English-specific dimension. Our alignment methodology builds empirical counterparts to these four scores from Bloomberg news each day.

Full-information benchmark and language-segmented processing. Let Σ represent the covariance matrix of s and $\lambda \equiv \text{Cov}(s, x)$. A hypothetical full-information, omnilingual trader who knows (Σ, λ) would optimally form a posterior mean for x as a linear combination of the four scores using Bayes' rule: $\widehat{x}^{\text{full}} = \omega^\top s$, where $\omega = (\omega_{AJ}, \omega_{AE}, \omega_{UJ}, \omega_{UE})^\top$ is the vector of full-information Bayesian weights. Each ω_k measures how much score k should contribute to forecasting x when all four scores are combined as efficiently as possible. We assume scores are signed so that higher scores are good news: $\omega \geq 0$.

Language frictions enter through *perceived* noise in non-native-language news. We say there is *language-segmented processing* when each investor type $i \in \{E, J\}$ works with a perceived covariance matrix $\Sigma_i \neq \Sigma$. As a baseline, we assume investors perceive non-native-language scores to be noisier than objectively warranted.[6] Intuitively, Japanese investors

[6] We assume investors agree to disagree about fixed perceived covariance matrices Σ_i and do not attempt to infer a language-neutral covariance matrix from prices. Our framework is thus close in spirit to models of disagreement over signal precision such as Scheinkman and Xiong (2003). One could imagine a richer dynamic environment in which investors gradually learn about the informativeness of foreign-language news and adjust Σ_i over time, which would weaken language segmentation. Our framework is a static reduced-form description of the resulting segmentation at a point in time; the effects we document can be interpreted as evidence that such learning, if it occurs, has not fully eliminated language-segmented processing.

inflate the noise in English scores, while foreign investors inflate the noise in Japanese scores. Given Σ_i, investors form posterior beliefs about x using Bayes' rule, and in this baseline, underweight non-native-language scores relative to the full-information benchmark.

Prices, returns, and a measure of inefficiency. With mean-variance preferences and Gaussian distributions, the equilibrium price can be written as $P = \beta^\top s$ where $\beta = (\beta_{AJ}, \beta_{AE}, \beta_{UJ}, \beta_{UE})^\top$ is a price weight vector. Each β_k is function of the belief weights that English and Japanese investors apply to each signal, the mass of investors q_J and q_E, and their risk-bearing capacities. The realized return tomorrow is $R \equiv x - P$.

Expected returns conditional on s equal $E[R \mid s] = \sum_k \gamma_k s_k$ for $k \in \{AJ, AE, UJ, UE\}$, where each slope γ_k is the gap between the full-information and price-implied weights, $\gamma_k = \omega_k - \beta_k$. We interpret γ_k as a measure of inefficiency. A value of $\gamma_k > 0$ means prices underweight score k relative to its full-information contribution. In contrast, in the full-information omnilingual benchmark, $\gamma_k = 0$ since $\beta = \omega$.

Our empirical strategy is to estimate the γ_k's and interpret their pattern. Specifically, we ask: (i) holding content fixed via alignment, does news in at least one language still forecast returns, indicating that prices underweight some content-matched signals across languages? and (ii) do unaligned, language-specific scores also forecast returns, revealing information investors miss in one language stream? The language-segmented processing framework formalizes these questions as the following predictions.

Prediction 1: Underreaction to unaligned news. *Suppose $\omega_k > 0$ for $k \in \{UE, UJ\}$. If investors perceive non-native-language scores to be noisier than an omnilingual investor, then $\gamma_k > 0$.*

The intuition is straightforward. For an informative unaligned signal such as s_{UE}, English-language investors put the full-information weight on it, while Japanese-language investors put a smaller weight because they treat it as noisier. The price weight on s_{UE} is a convex combination of these two weights and falls below ω_{UE}, leading the price to underreact. Evidence for this prediction would suggest that the predictive power of news in different languages is at least partly explained by differences in coverage.

Prediction 2: Underreaction to aligned news. *Suppose* $\omega_k > 0$ *for* $k \in \{AE, AJ\}$. *If investors perceive non-native-language scores to be noisier than an omnilingual investor, then* $\gamma_{AE} + \gamma_{AJ} > 0$.

The intuition for Prediction 2 is that the price underweights the combined information about the dual-language dimension of a stock's payoff. Both English and Japanese investors see s_{AE} and s_{AJ}, but each type discounts the non-native-language signal and thus puts too little total weight on the pair (s_{AE}, s_{AJ}) relative to an omnilingual investor who would optimally combine both. Since no one in the market fully uses the combined information, the price underreacts relative to the full-information benchmark: $\beta_{AE} + \beta_{AJ} < \omega_{AE} + \omega_{AJ}$, implying $\gamma_{AE} + \gamma_{AJ} > 0$. Evidence for Prediction 2 would suggest that the predictive power of news in different languages cannot be explained purely by unaligned coverage. It persists even when we hold event content fixed and include AE and AJ scores together.[7]

This baseline assumes that investors inflate the perceived noise of non-native-language signals, but relaxing this assumption and allowing more general misperception of signals by language generates richer behavior. For example, if some investors instead overestimate the precision of certain news and thus deflate the noise, e.g., because Japanese news is perceived as conferring an informational edge, then prices can overweight those details and reverse quickly. We discuss mixed evidence of this possibility in Sections 3.4-3.5.

3.2 Alignment statistics and correlations

Table 5 summarizes the properties of our aligned and unaligned sentence embeddings for Japanese stocks. Panel A reports cosine similarities between English and Japanese sentence embeddings. Aligned sentences have an average cosine similarity of 0.81, versus 0.48 for unaligned sentences and 0.76 for all sentences. This confirms that the alignment procedure identifies sentence pairs that are much closer in semantic space than unmatched sentences.

[7]Note that the framework does not pin down the signs of individual γ_{AE}, γ_{AJ} beyond what is stated in Prediction 1. The framework disciplines the sum of these forces for the pair (AE, AJ) through the common-news structure and the language friction but does not impose enough structure to sign γ_{AE}, γ_{AJ} individually. Either could be negative. For example, we could have $\gamma_{AJ} < 0$ in Japanese markets if most price setters are Japanese and think English signals are extremely noisy, leading prices to overweight s_{AJ} and underweight s_{AE} relative to the full-information benchmark.

Panel B reports summary statistics for the return scores. While roughly comparable to those from Table 2, differences can emerge because Table 5 has a smaller sample size (a stock needs to have both JP and EN news for our alignment process to work) and because the alignment process focuses on article bodies instead of the head+body.

Panel C reports correlations between the resulting return scores. The aligned English and Japanese scores (AE and AJ) are more correlated with each other (0.65) than they are with their own-language unaligned scores (0.40 for AE/UE, 0.42 for AJ/UJ). Correlations between aligned scores and cross-language unaligned scores are also lower (0.34 for AE/UJ, 0.46 for AJ/UE). These statistics support the premise of the conceptual framework: AE and AJ measure the shared core A, while UE and UJ capture additional information that appears only in one language.

3.3 Unaligned and aligned return predictability

We now study Predictions 1 and 2 for Japanese stocks. Table 6 reports daily cross-sectional regressions of open–to–close returns on aligned and unaligned scores, estimated only on stock–days with news in both languages:

$$\begin{aligned}\mathrm{Ret}_{i,t}^{O\to C} = a_t &+ b_0\,\mathbb{1}_{i,t}^{AE} + b_1\,\mathbb{1}_{i,t}^{AJ} + b_2\,\mathbb{1}_{i,t}^{UE} + b_3\,\mathbb{1}_{i,t}^{UJ} \\ &+ b_4\,\mathbb{1}_{i,t}^{AE}\times\widehat{R}_{i,t}^{AE,pre} + b_5\,\mathbb{1}_{i,t}^{AJ}\times\widehat{R}_{i,t}^{AJ,pre} + b_6\,\mathbb{1}_{i,t}^{UE}\times\widehat{R}_{i,t}^{UE,pre} + b_7\,\mathbb{1}_{i,t}^{UJ}\times\widehat{R}_{i,t}^{UJ,pre} \\ &+ \Gamma' X_{i,t} + \varepsilon_{i,t},\end{aligned}$$

where $\mathbb{1}_{i,t}^{k}$ indicates the presence of at least one sentence of type $k \in \{AE, AJ, UE, UJ\}$, $\widehat{R}_{i,t}^{k,pre}$ is the corresponding return score, and $X_{i,t}$ includes log market capitalization and Dimson beta. The indicator variables account for the possibility that, even within our sample of stock-days with news in both languages, all or no sentences from one language may align with sentences from the other language. All specifications include date fixed effects (a_t) and we cluster standard errors by firm and date. Three main facts emerge.

First, unaligned English news (UE) robustly predicts returns with a positive slope in all columns (Prediction 1). The gradual inclusion of all four scores and controls attenuates the slope on $UE \times \widehat{R}^{UE}$ from 0.36 (column (1)) to 0.14 (column (5)), but UE retains predictive

power throughout. A large, positive UE coefficient is natural in the language-segmented processing framework, as domestic Japanese investors who dominate the market may miss or underweight information available in the foreign language.

Second, aligned scores in both languages (AE and AJ) have large and statistically significant positive slopes (Prediction 2). In columns (1)–(2), the coefficients on $AE \times \widehat{R}^{AE}$ and $AJ \times \widehat{R}^{AJ}$ are roughly 0.53 and 0.61, respectively, and attenuate to 0.22 with the gradual inclusion of all four scores and control variables. The near-identical coefficients in column (5) indicate that prices underreact to English and Japanese aligned news in nearly identical fashion. In all cases, the coefficients remain economically and statistically greater than zero. Thus, even after aligning sentences across languages and isolating the shared content, aligned English and Japanese signals continue to predict subsequent Japanese returns.

Third, the results for unaligned Japanese news (UJ) are mixed, with near-zero coefficients in columns (1)-(4) and a negative, marginally significant coefficient in column (5). The negative slope on UJ hints at overreaction and investor overconfidence in details surfaced specifically by domestic Japanese news, and can exist even when slopes on AJ are positive.[8]

Taken together, these findings suggest that the return predictability of Japanese scores is primarily associated with sentences that share content with English, rather than with Japanese-only text. Conversely, the return predictability of English scores is due to a mix of sentences that appear only in English and those that are shared between both languages.

3.4 Heterogeneity by size and ownership

Table 7 studies how the return predictability of aligned and unaligned scores varies across market capitalization and foreign ownership share.[9] The results suggest three observations.

First, small and domestically owned firms have stock prices that underreact to aligned news. The slopes on the aligned scores AE and AJ are substantially larger for small firms than for large firms (columns (1)–(3)) and for domestically owned firms than for firms with high foreign ownership (columns (4)–(6)). Column (7) reinforces the small stock effect.

[8] Intuitively, overreaction to Japanese-only details need not imply overreaction to aligned Japanese news because aligned English coverage provides corroborating information that can discipline pricing. The Internet Appendix provides details.

[9] The foreign ownership data is retrieved from Astra Manager.

Figure 7 shows that long-short strategies formed on aligned scores are highly profitable within small stocks and low foreign ownership stocks, but are not significantly profitable within larger or higher ownership stocks.

Second, the evidence on unaligned English news in the sub-sample of small stocks is consistent with that of Table 6, although statistically weaker. Although only marginally significant, the magnitude of the slope on UE is largest when looking at small stocks in column (1) and is over twice the slope on larger stocks in column (2).

Third, the evidence on unaligned Japanese news suggests that prices of large, foreign-owned stocks can overreact. The estimates in columns (3) and (6) show that the slopes on UJ are negative for larger and higher-foreign-ownership stocks but are much closer to zero for smaller and more-domestically-owned stocks.[10]

Taking all findings together, the pattern supports language-segmented processing, which predicts stronger effects where local investors dominate and cross-language arbitrage is weakest: small, domestically-owned Japanese firms. Prices underreact to the aligned signals in these stocks and not in larger, more foreign-owned firms, consistent with language frictions manifesting themselves where stocks face more limits-to-arbitrage. The potential overreaction to unaligned Japanese news among large, foreign-owned stocks is consistent with additional layers of language-segmented processing under the relaxed assumptions described at the end of Section 3.1.

3.5 Readership as a direct attention proxy

Table 8 uses readership as a direct proxy for attention to each language stream. For each stock–day, we construct indicators $\mathbb{1}_{i,t}^{RE}$ and $\mathbb{1}_{i,t}^{RJ}$ that equal one if at least one article in that language has at least 25 Bloomberg readers. We then estimate regressions of returns on our scores interacted with these readership indicators.

In columns (1)-(2), we study language-level scores. The slopes on $\widehat{R}^{EN}$ and $\widehat{R}^{JP}$ are positive, but their interactions with high readership in English and Japanese (respectively) are negative. These results suggest that prices underreact to news in each language, but the

[10] Our inclusion of time effects means the slopes in columns (1) and (2) do not strictly sum to those in (3), and similarly for columns (4)-(6).

underreaction dissipates when readership of news in that language is high.

In columns (3)-(4), we study our unaligned and aligned scores. For unaligned English scores, the underreaction persists even when readership is high but suggests some attentuation through negative point estimates of the $RE \times UE \times \hat{R}^{UE}$ interaction. This result indicates that unaligned English news is particularly hard for some investors to pick up even when they access English news.

Among aligned scores, greater readership reduces underreaction significantly. The slopes on AE and AJ are positive when readerships of English and Japanese articles (respectively) are low, but moves toward zero when readerships are high. Figure 7 shows that long-short strategies formed on aligned scores are profitable within low-readership stock-days but less profitable on high-readership stock-days.

For unaligned Japanese scores, prices overreact when readership is high. Specifically, the slopes are zero when Japanese article readership is low, but become strongly negative when that readership is high. This pattern is consistent with that from Table 7 with investors overweighting Japanese-only details in certain situations.

Taken together, these results reinforce a language-based interpretation: pricing of news in a given language is more inefficient when article readership in that language is low, and closer to efficient when readership is high. The exception is in unaligned Japanese news, when prices tend to overreact when readership is high.

3.6 U.S. Markets

Table 9 repeats the analysis for U.S. stocks and news. Panel A reports correlations for our aligned/unaligned scores and exhibits a similar pattern to Table 5: aligned scores across languages tend to correlate more than with own-language unaligned scores. Panel B studies the aligned/unaligned cross-sectional return forecasting regression for U.S. stocks.

The table shows that aligned English news (AE) has large and statistically significant slopes for U.S. stocks. In the full specification with all four scores and controls (column (5)), the coefficient on $\mathbb{1}^{AE} \times \widehat{R}^{AE}$ is 0.38 and remains significantly greater than zero. This is consistent with underreaction to the main English–language coverage of U.S. firms. By contrast, aligned Japanese news (AJ) has a near-zero point estimate once AE is included, po-

tentially indicating that aligned Japanese news is largely not objectively informative beyond AE news. For the unaligned components, coefficients are generally small and imprecisely estimated once controls are added, suggesting that most of the return-forecasting content in this setting is concentrated in the aligned English component.

Overall, the U.S. results highlight an asymmetry relative to Japan. For Japanese stocks, both aligned scores are strongly predictive and much of the incremental predictability beyond aligned coverage comes from English unaligned news. For U.S. stocks, predictability is concentrated in aligned English news, while Japanese aligned and English unaligned scores play a smaller role. This asymmetry contrasts with the results in Table 4, where both English and Japanese news scores predict returns and suggests that the aligned English news captures a significant piece of predictive power across both languages.

4 Robustness and Discussion

4.1 Robustness

Table 10 reports a series of robustness checks that address several concerns: (i) potential look–ahead bias from using pre–trained language models, (ii) sensitivity to the alignment procedure, (iii) dependence on a particular choice of sentence–embedding model, and (iv) dependence on the window of news.

Post–2020 sample and look–ahead bias. A natural concern is that LaBSE, the multilingual model we use to build sentence embeddings, was trained on text available prior to its public release in 2020. The embeddings could in principle encode information about future news, inflating predictability through look-ahead bias. Panel A of Table 10 addresses this concern by re–estimating the language–level return forecasting regressions using only data after June 1, 2020, so that all news in this subsample is strictly out–of–sample relative to LaBSE's training period. The coefficients on the English and Japanese return scores remain large and statistically significant, with magnitudes similar to (if somewhat noisier than) the full–sample estimates.

Panel B repeats the aligned/unaligned regressions for Japanese stocks on the post–2020

subsample. The main message is unchanged in that unaligned and aligned scores continue to predict returns post-2020 across all but the most stringent specification. The slopes for UE and UJ are not statistically significant in column (5), although the effective sample size is less than half of that in Table 6 and magnitudes are nevertheless comparable. Indeed, the slope on UJ is more consistently positive than in Table 6. These changes indicate our earlier results are not driven by look-ahead bias.

Alternative alignment configuration. Panel C implements a more stringent "mutual top" alignment that is stronger than our baseline "global top-mass" alignment method. Whereas our baseline method keeps only those sentence pairs whose optimal-transport probability mass lies in the top slice of the transport plan in either direction, treating them as the strongest overall matches, "mutual top" adds the requirement that a matched sentence pair to be simultaneously high-mass and mutually top-ranked in both directions.[11] This reduces the number of aligned sentences and shifts more text into the unaligned buckets, providing a conservative check on our baseline alignment procedure.

Re–estimating the aligned/unaligned regressions for Japanese stocks with this stricter alignment yields coefficients that are broadly similar to those in Table 6. Thus, our results are not sensitive to this alternative alignment configuration.

Alternative embedding models. Panel D recomputes the language–level return scores for Japanese stocks using two alternative multilingual sentence–embedding models, GTE and E5–multilingual, instead of LaBSE. Panel E does the same for U.S. stocks, providing analogues to the baseline language–level regressions for Japan and the U.S. from Table 4. Across all four panels, the English and Japanese return scores constructed from GTE and E5 continue to strongly predict next–day returns. The magnitudes and relative ordering of the coefficients are broadly similar to the LaBSE baseline, with both languages exhibiting sizable predictive power and English remaining the dominant predictor for U.S. stocks. Taken together, these checks suggest that our language-level predictability is robust to the choice

[11] Specifically, mutual top alignment is a 1-to-1 mapping. Within each row i in γ, we identify the maximum element $\gamma_{ij_{max}}$. If $\gamma_{ij_{max}}$ falls within the top 5% of column j_{max}, we mark a forward alignment, producing the binary forward alignment matrix $A \in \{0,1\}^{n\times m}$. The backward alignment matrix $B \in \{0,1\}^{m\times n}$ is computed analogously.

of sentence–embedding model.[12]

4.2 Practical Implications

Our results suggest several concrete ways that practitioners can use multilingual news and large language models.

1. Use LLMs to turn raw news into return forecasts. Multilingual language models can be turned into return forecasting signals, reinforcing results from Chen, Kelly, and Xiu (2022). In both Japan and the U.S., simple linear models trained on embeddings from LaBSE, GTE, or E5 applied to Bloomberg headlines and body text produce economically large, positive alphas at daily horizons, with text from article bodies (not just headlines) providing meaningful information. This process skips bag–of–words engineering, sacrificing a level of interpretability, but generates significant alpha based on clean news feeds.

2. Be the omnilingual trader. There is value in systematically processing news reported in other languages. For Japanese stocks, both Japanese and English news streams independently forecast returns, and processing both news streams captures news specific to each language while also more-efficiently processing the content that is shared across languages. Allowing both local– and foreign–language signals to contribute to trading strategies, even when they appear to be covering the same event, improves performance.

3. Inform strategies by finding what is similar and different across news feeds. Our alignment procedure separates shared coverage from language-specific content. When a strategy takes a position based on Japanese and English news, alignment can attribute the signal to content reported in both languages, to information appearing in only one, or to both. For Japanese stocks, we find that both aligned content and English-only sentences contribute to predictability, while Japanese-only sentences add little. This decomposition provides explainability and interpretability: rather than relying on a simple combination of signals from Japanese and English news, a trader can study whether a position reflects news

[12] The GTE and E5 models do not currently produce sufficiently differentiated sentence embeddings to generate a meaningful alignment process.

that both investor groups have seen or information that domestic Japanese investors may have missed, as well as the specific sentences in each group. More broadly, our alignment method offers a practical way to study how to efficiently combine information across different news feeds to inform trading.

4. Tilt toward segments where language and attention frictions are strongest. In Japan, the predictive power of aligned scores is stronger in small–cap and more domestically owned firms than in large, internationally held firms. This suggests that strategies that place more weight on multilingual news signals in these segments can generate greater alpha. Moreover, return predictability for a given language's return score is stronger when articles in that language are less read and shrinks toward zero (and sometimes overshoots) when those articles attract more readership. In practice, this points to a simple refinement: scale the aggressiveness of language-based signals based on how widely the underlying stories *in that same language* are being read.

To demonstrate, we calculate the performance of strategies that sort on aligned and unaligned signals. We average the two aligned signals AE and AJ to form a composite "A" signal, capturing content reported in both languages. We then sort Japanese stocks each day into quartiles of composite aligned news and unaligned English news.

Sorting on both dimensions generates substantial alpha. Figure 8 shows the average returns of each portfolio: moving from low- to high-A portfolios increases returns, as does moving from low- to high-UE portfolios. Moving along the diagonal, from low-A/low-UE to high-A/high-UE portfolios, generates the largest returns. Table 11 reports that the alpha from a long-short strategy across this diagonal is nearly 50 points per day. This strategy generates alpha greater than those from single-language strategies in Table 3 and captures nearly all of the alpha from double-sort strategies based on just languages, with the key difference that our decomposition reveals what is driving each position. A stock in the high-A/high-UE corner, for instance, has positive shared content and English-only coverage that domestic investors may have missed. A stock in the high-A/low-UE corner reflects news both groups have seen. On any given day, the trader can examine the specific sentences in

each bucket.

5 Conclusion

Our evidence indicates that markets are not "omnilingual": investors not only miss news reported in only one language, but also do not efficiently use the same news reported across multiple languages to sharpen their views. We show this using multilingual large language models on fifteen years of Japanese- and English-language Bloomberg news for Japanese and U.S. equities. By aligning sentences across languages, we separate overlapping content from language-specific coverage and build return-forecasting signals from each group. We find that both unaligned and aligned signals predict returns, with stronger effects where local investors dominate and readership is low. These patterns suggest that language boundaries remain a distinct friction in price discovery.

References

Ball, R., and P. Brown. 1968. An empirical evaluation of accounting income numbers. *Journal of Accounting Research* 6:159–78.

Bernard, V. L., and J. K. Thomas. 1989. Post-earnings-announcement drift: Delayed price response or risk premium? *Journal of Accounting Research* 27:1–36.

Bybee, L., B. Kelly, A. Manela, and D. Xiu. 2024. Business news and business cycles. *The Journal of Finance* 79:3105–47. doi:https://doi.org/10.1111/jofi.13377.

Caffarelli, L. A., M. Feldman, and R. J. McCann. 2002. Constructing optimal maps for monge's transport problem as a limit of strictly convex costs. *Journal of the American Mathematical Society* 15:1–26.

Chen, J., G. Tang, G. Zhou, and W. Zhu. 2025. ChatGPT and deepseek: Can they predict the stock market and macroeconomy? Working paper.

Chen, Y., B. T. Kelly, and D. Xiu. 2022. Expected returns and large language models. *Available at SSRN 4416687* .

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–86. Minneapolis, Minnesota: Association for Computational Linguistics.

Dimson, E. 1979. Risk measurement when shares are subject to infrequent trading. *Journal of Financial Economics* 7:197–226.

Dou, Z.-Y., and G. Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In P. Merlo, J. Tiedemann, and R. Tsarfaty, eds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2112–28. Online: Association for Computational Linguistics.

Du, Z., A. G. Huang, R. Wermers, and W. Wu. 2022. Language and domain specificity: A chinese financial sentiment dictionary. *Review of Finance* 26:673–719.

Durand, R. B., J. Khuu, and L. A. Smales. 2023. Lost in translation. when sentiment metrics for one market are derived from two different languages. *Journal of Behavioral and Experimental Finance* 39:100825–.

Engelberg, J., A. Manela, W. Mullins, and L. Vulicevic. 2025. Entity neutering. Working paper.

Feng, F., Y. Yang, D. Cer, N. Arivazhagan, and W. Wang. 2022. Language-agnostic BERT sentence embedding. In S. Muresan, P. Nakov, and A. Villavicencio, eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 878–91. Dublin, Ireland: Association for Computational Linguistics.

Grinblatt, M., and M. Keloharju. 2001. How distance, language, and culture influence stockholdings and trades. *The Journal of Finance* 56:1053–73.

Gu, S., B. Kelly, and D. Xiu. 2020. Empirical asset pricing via machine learning. *The Review of Financial Studies* 33:2223–73.

He, S., L. Lv, A. Manela, and J. Wu. 2025. Chronologically consistent large language models. Working paper.

Hong, H., and J. C. Stein. 1999. A unified theory of underreaction, momentum trading, and overreaction in asset markets. *The Journal of Finance* 54:2143–84.

Honnibal, M., I. Montani, S. Van Landeghem, and A. Boyd. 2020. spacy: Industrial-strength natural language processing in python. *Zenodo* doi:10.5281/zenodo.1212303.

Kantorovich, L. V. 1942. On the translocation of mass. *Doklady Akademii Nauk SSSR* 37:199–201. In Russian: "O peremeschenii mass".

Lang, T., L. Stice-Lawrence, Y. T. F. Wong, and T. Wong. 2025. Differential communication and local information advantage: Revelations from translation differences. Working paper.

Levy, B. 2024. Caution ahead: Numerical reasoning and look-ahead bias in AI models. Working paper.

Lopez-Lira, A., and Y. Tang. 2025. Can chatGPT forecast stock price movements? Return predictability and large language models. ArXiv working paper 2304.07619.

Loughran, T., and B. McDonald. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance* 66:35–65.

Manela, A., and A. Moreira. 2017. News implied volatility and disaster concerns. *Journal of Financial Economics* 123:137–62.

Martineau, C. 2022. Rest in peace post-earnings announcement drift. *Critical Finance Review* 11:613–46.

Monge, G. 1781. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris* .

Peyré, G., and M. Cuturi. 2020. Computational optimal transport. ArXiv working paper 1803.00567.

Rahimikia, E., and F. Drinkall. 2024. Re(visiting) large language models in finance. SSRN Working Paper #4963618.

Sarkar, S., and K. Vafa. 2024. Lookahead bias in pretrained language models. SSRN Working Paper #4754678.

Scheinkman, J., and W. Xiong. 2003. Overconfidence and speculative bubbles. *Journal of Political Economy* 111:1183–220.

Sinkhorn, R. 1964. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The Annals of Mathematical Statistics* 35:876–9.

Swanson, K., L. Yu, and T. Lei. 2020. Rationalizing text matching: Learning sparse alignments via optimal transport. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5609–26. Online: Association for Computational Linguistics.

Tetlock, P. C. 2007. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance* 62:1139–68.

Wang, C., and Y. Qiu. 2025. The sparse-plus-low-rank quasi-newton method for entropic-regularized optimal transport. In *Forty-second International Conference on Machine Learning*.

Wang, L., N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei. 2024. Multilingual e5 text embeddings: A technical report.

Zhang, X., Y. Zhang, D. Long, W. Xie, Z. Dai, J. Tang, H. Lin, B. Yang, P. Xie, F. Huang, M. Zhang, W. Li, and M. Zhang. 2024. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In F. Dernoncourt, D. Preoţiuc-Pietro, and A. Shimorina, eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 1393–412. Miami, Florida, US: Association for Computational Linguistics.

Table 1: Sample Summary Statistics

Panel A of this table reports summary statistics for the number of unique Japanese and US stocks in our sample. Panels B and C report the average number of news per year for large–and small–cap Japanese and US stocks, respectively. For Japanese stocks, we create market capitalization breakpoints using quintile cutoffs for the largest 1,500 stocks. We apply the breakpoint cutoffs to all stocks using the previous month's capitalization breakpoints. We define large (small) stocks as those in the top (bottom two) quintiles. We do the same classification for US stocks using the Fama-French market capitalization quintile breakpoints.

A. Sample Stocks Summary

		Yearly Average Number of Stocks		
	Number of Stocks	All	Large Cap	Small Cap
Japanese Stocks	3,797	3,060	868	2,540
US Stocks	7,362	3,899	1,202	3,222

B. Japanese Stocks

		Yearly Average Number of News		
	Number of News	All	Large Cap	Small Cap
Japanese News	943,085	58,670	32,676	25,994
English News	1,159,771	71,677	43,358	28,319

C. US Stocks

		Yearly Average Number of News		
	Number of News	All	Large Cap	Small Cap
Japanese News	193,593	12,058	9,294	2,764
English News	5,462,654	336,514	194,722	141,792

Table 2: Return Score Summary Statistics

This table reports summary statistics for the predicted returns, $\widehat{R}_{i,t}^{\mathrm{l,\ pre}}$, constructed from text embeddings generated in each language l using three models: LaBSE, GTE, and E5. For brevity, the table reports summary statistics for articles in head+body format. Panel A reports the mean, standard deviation, median, and the 25th and 75th percentiles of $\widehat{R}_{i,t}^{\mathrm{l,\ pre}}$ Japanese stocks. Panel B provides the corresponding statistics for US stocks. The predicted returns are computed from 2014 to 2024.

A. Japanese stocks

		Mean	Std	Median	25^{th}	75^{th}
LaBSE	EN	-0.10%	0.17%	-0.06%	-0.20%	0.00%
	JP	-0.12%	0.20%	-0.09%	-0.24%	0.00%
GTE	EN	-0.10%	0.16%	-0.07%	-0.19%	0.00%
	JP	-0.11%	0.16%	-0.09%	-0.21%	0.00%
E5	EN	-0.11%	0.16%	-0.08%	-0.21%	0.00%
	JP	-0.12%	0.19%	-0.09%	-0.23%	0.00%

B. U.S. stocks

		Mean	Std	Median	25^{th}	75^{th}
LaBSE	EN	-0.05%	0.22%	-0.04%	-0.18%	0.09%
	JP	-0.00%	0.05%	0.00%	0.00%	0.00%
GTE	EN	-0.05%	0.22%	-0.03%	-0.17%	0.08%
	JP	-0.00%	0.05%	0.00%	0.00%	0.00%
E5	EN	-0.06%	0.23%	-0.04%	-0.19%	0.09%
	JP	-0.00%	0.05%	0.00%	0.00%	0.00%

Table 3: Model Performance

Panel A reports the best annualized Sharpe ratios for Japanese and US stocks using trading strategies formed by sorting stocks on predicted returns, $\widehat{R}^{l}_{i,t}$. The sample is restricted to days with at least 20 news-covered stocks. The long–short return is defined as $LS^{l}_{t} = L^{l}_{t} - S^{l}_{t}$, where L^{l}_{t} is the average return of stocks in the top 25% of $\widehat{R}^{\mathrm{l,\ pre}}_{i,t}$, and S^{l}_{t} is the average return of those in the bottom 25%. For each model (LaBSE, GTE, E5), the first row reports the best annualized Sharpe ratio, and the second row lists the optimal configuration: language (Japanese or English) and text type (headline, body, or headbody). Panels B-C report CAPM and Fama–French five-factor+momentum alphas. Returns and alphas are reported as daily percentage points. ***, **, * corresponds to $p < .01$, $p < 0.05$, and $p < .1$, respectively.

A. Sharpe Ratios

Model	Japan	US
LaBSE	5.29	5.34
	Japanese (headbody)	English (headline)
GTE	4.69	6.00
	English (headbody)	English (headbody)
E5	5.03	6.26
	Japanese (headline)	English (headline)

B. CAPM Alphas (Headbody Japanese News)

	LaBSE			GTE			E5		
	Short	Long	Long-Short	Short	Long	Long-Short	Short	Long	Long-Short
Average Return	-0.341***	-0.023	0.318***	-0.296***	-0.030*	0.266***	-0.316***	-0.019	0.297***
Standard Deviation	1.100	0.925	0.954	1.099	0.888	0.933	1.139	0.879	0.979
CAPM Alpha	-0.354***	-0.038**	0.316***	-0.310***	-0.045***	0.264***	-0.330***	-0.035**	0.295***
FF5+mom Alpha	-0.350***	-0.036**	0.314***	-0.307***	-0.043***	0.264***	-0.325***	-0.034**	0.291***

C. CAPM Alphas (Headbody English News)

	LaBSE			GTE			E5		
	Short	Long	Long-Short	Short	Long	Long-Short	Short	Long	Long-Short
Average Return	-0.250***	-0.011	0.239***	-0.262***	0.003	0.265***	-0.261***	-0.004	0.257***
Standard Deviation	1.072	0.912	0.904	1.066	0.887	0.897	1.088	0.867	0.919
CAPM Alpha	-0.263***	-0.025*	0.237***	-0.274***	-0.011	0.263***	-0.274***	-0.018	0.255***
FF5+mom Alpha	-0.261***	-0.025*	0.236***	-0.273***	-0.010	0.263***	-0.271***	-0.017	0.253***

Table 4: Cross-sectional Regressions

This table reports the following regressions:

$$R_{i,t}^{O\to C} = a_t + b_0 \widehat{R}_{i,t}^{\text{EN, pre}} + b_1 \widehat{R}_{i,t}^{\text{JP, pre}} + \text{Control}_{i,t} + \varepsilon_{i,t},$$

where $R_{i,t}^{O\to C} = \frac{\text{Close}_t - \text{Open}_t}{\text{Open}_t}$ is the same-day open–close return, and $\widehat{R}_{i,t}^{\text{l, pre}}$ is the corresponding predicted return score. Controls include the firm's log market capitalization on $t-1$ and its Dimson beta. All specifications include time fixed effects. Column (4) adds the control variables. Standard errors are two-way clustered by firm and date. ***, **, * corresponds to $p < .01$, $p < 0.05$, and $p < .1$, respectively. Panel A reports estimates for the Japanese stocks using LaBSE embedding from the headbody text. Panel B reports the corresponding results for the US stocks.

A. Japanese stocks

	(1)	(2)	(3)	(4)
$\widehat{R}^{\text{EN, pre}}$	0.617***		0.338***	0.166***
	(0.04)		(0.04)	(0.04)
$\widehat{R}^{\text{JP, pre}}$		0.697***	0.581***	0.469***
		(0.04)	(0.05)	(0.04)
Log(mcap)				0.039***
				(0.00)
Beta				-0.038*
				(0.02)
N	340,309	340,309	340,309	331,290
T	2,627	2,627	2,627	2,627
R^2 (%)	0.12	0.21	0.24	0.34

B. U.S. stocks

	(1)	(2)	(3)	(4)
$\widehat{R}^{\text{EN, pre}}$	0.646***		0.640***	0.508***
	(0.04)		(0.04)	(0.03)
$\widehat{R}^{\text{JP, pre}}$		0.510***	0.326***	0.404***
		(0.08)	(0.08)	(0.08)
Log(mcap)				0.048***
				(0.01)
Beta				-0.085***
				(0.02)
N	934,742	934,742	934,742	887,596
T	2,705	2,705	2,705	2,705
R^2 (%)	0.09	0.00	0.09	0.17

Table 5: Alignment Statistics for Japanese Stocks

Panel A reports summary statistics for the cosine similarities of English–Japanese article pairs, computed using aligned, unaligned, and full embeddings over the 2008–2024 sample period. Panel B reports summary statistics for the return scores. Panel C presents the correlation matrix of predicted returns, $\widehat{R}_{i,t}^{\mathrm{l,\ pre}}$, where $l \in \{AE, UE, EN, AJ, UJ, JP\}$, where the EN, JP superscripts denote predicted returns based on full embeddings and AE, UE, AJ, UJ denote embeddings from Aligned-English, Unaligned-English, Aligned-Japanese, and Unaligned-Japanese groups of sentences.

A. Cosine Similarities

Type	Mean	Std	5^{th}	50^{th}	95^{th}
aligned	0.81	0.05	0.70	0.82	0.88
unaligned	0.48	0.18	0.19	0.47	0.77
full	0.76	0.08	0.60	0.77	0.86

B. Return Score Summary Statistics

Type	Mean	Std	Median	25%	75%
AE	-0.17%	0.22%	-0.17%	-0.31%	-0.04%
AJ	-0.18%	0.22%	-0.18%	-0.32%	-0.04%
UE	-0.15%	0.19%	-0.13%	-0.26%	-0.03%
UJ	-0.15%	0.13%	-0.15%	-0.23%	-0.07%

C. Return Score Correlation

	$\widehat{R}^{\mathrm{AE}}$	$\widehat{R}^{\mathrm{UE}}$	$\widehat{R}^{\mathrm{EN}}$	$\widehat{R}^{\mathrm{AJ}}$	$\widehat{R}^{\mathrm{UJ}}$	$\widehat{R}^{\mathrm{JP}}$
$\widehat{R}^{\mathrm{AE}}$	1.00	0.40	0.65	0.65	0.34	0.56
$\widehat{R}^{\mathrm{UE}}$	0.40	1.00	0.79	0.46	0.45	0.47
$\widehat{R}^{\mathrm{EN}}$	0.65	0.79	1.00	0.58	0.39	0.58
$\widehat{R}^{\mathrm{AJ}}$	0.65	0.46	0.58	1.00	0.42	0.74
$\widehat{R}^{\mathrm{UJ}}$	0.34	0.45	0.39	0.42	1.00	0.62
$\widehat{R}^{\mathrm{JP}}$	0.56	0.47	0.58	0.74	0.62	1.00

Table 6: Aligned and Unaligned Regressions for Japanese Stocks

This table reports the following regressions:

$$R_{i,t}^{O\to C} = a_t + b_0 \mathbb{1}_{i,t}^{\text{AE}} + b_1 \mathbb{1}_{i,t}^{\text{AJ}} + b_2 \mathbb{1}_{i,t}^{\text{UE}} + b_3 \mathbb{1}_{i,t}^{\text{UJ}} + b_4 \mathbb{1}_{i,t}^{\text{AE}} \times \widehat{R}_{i,t}^{\text{AE, pre}} + b_5 \mathbb{1}_{i,t}^{\text{AJ}} \times \widehat{R}_{i,t}^{\text{AJ, pre}} + b_6 \mathbb{1}_{i,t}^{\text{UE}} \times \widehat{R}_{i,t}^{\text{UE, pre}} + b_7 \mathbb{1}_{i,t}^{\text{UJ}} \times \widehat{R}_{i,t}^{\text{UJ, pre}} + \text{Control}_{i,t} + \varepsilon_{i,t},$$

where $R_{i,t}^{O\to C}$ is the open–close return, $\mathbb{1}_{i,t}^{l}$ indicates if firm i has aligned or unaligned news sentences in $l \in \{AE, AJ, UE, UJ\}$ on day t, and $\widehat{R}_{i,t}^{\text{l, pre}}$ is the corresponding predicted return score. Controls include the firm's log market capitalization on $t-1$ and its Dimson beta. All specifications include time fixed effects. Column (1) reports the estimates using AE and UE, Column (2) reports the estimates using AJ and UJ. Columns (3)–(5) include both English and Japanese news. $\mathbb{1}_{i,t}^{\text{AE}}$ is the omitted category. Column (5) adds the control variables to present the full specification. Standard errors are two-way clustered by firm and date. ***, **, * corresponds to $p < .01$, $p < 0.05$, and $p < .1$, respectively.

	(1)	(2)	(3)	(4)	(5)
$\mathbb{1}^{\text{AE}}$	-0.041				
	(0.03)				
$\mathbb{1}^{\text{AJ}}$		-0.064**		-0.033	0.027
		(0.03)		(0.03)	(0.03)
$\mathbb{1}^{\text{UE}}$	0.011			0.031	0.008
	(0.05)			(0.05)	(0.05)
$\mathbb{1}^{\text{UJ}}$		0.004		-0.018	-0.075***
		(0.02)		(0.02)	(0.03)
$\mathbb{1}^{\text{AE}} \times \widehat{R}^{\text{AE, pre}}$	0.528***		0.350***	0.348***	0.222***
	(0.05)		(0.05)	(0.05)	(0.05)
$\mathbb{1}^{\text{AJ}} \times \widehat{R}^{\text{AJ, pre}}$		0.609***	0.332***	0.332***	0.221***
		(0.05)	(0.05)	(0.05)	(0.05)
$\mathbb{1}^{\text{UE}} \times \widehat{R}^{\text{UE, pre}}$	0.361***		0.267***	0.271***	0.141**
	(0.06)		(0.06)	(0.06)	(0.06)
$\mathbb{1}^{\text{UJ}} \times \widehat{R}^{\text{UJ, pre}}$		0.096	0.015	-0.009	-0.166*
		(0.09)	(0.08)	(0.09)	(0.09)
Log(mcap)					0.056***
					(0.01)
Beta					-0.031
					(0.03)
N	177,995	177,995	177,995	177,995	172,959
T	2,624	2,624	2,624	2,624	2,624
R^2 (%)	0.19	0.17	0.21	0.21	0.31

Table 7: Heterogeneity Regression for Japanese Stocks

This table reports the following regressions:

$$R_{i,t}^{O\to C} = a_t + b_0 \mathbb{1}_{i,t}^{\text{AE}} \times \widehat{R}_{i,t}^{\text{AE, pre}} + b_1 \mathbb{1}_{i,t}^{\text{AJ}} \times \widehat{R}_{i,t}^{\text{AJ, pre}} + b_2 \mathbb{1}_{i,t}^{\text{UE}} \times \widehat{R}_{i,t}^{\text{UE},pre} + b_3 \mathbb{1}_{i,t}^{\text{UJ}} \times \widehat{R}_{i,t}^{\text{UJ, pre}} + b_4 \mathbb{1}_{i,t}^{h} \times \mathbb{1}_{i,t}^{\text{AE}} \times \widehat{R}_{i,t}^{\text{AE, pre}} + b_5 \mathbb{1}_{i,t}^{h} \times \mathbb{1}_{i,t}^{\text{AJ}} \times \widehat{R}_{i,t}^{\text{AJ, pre}} + b_6 \mathbb{1}_{i,t}^{h} \times \mathbb{1}_{i,t}^{\text{UE}} \times \widehat{R}_{i,t}^{\text{UE},pre} + b_7 \mathbb{1}_{i,t}^{h} \times \mathbb{1}_{i,t}^{\text{UJ}} \times \widehat{R}_{i,t}^{\text{UJ, pre}} + \text{Control}_{i,t} + \varepsilon_{i,t},$$

where $R_{i,t}^{O\to C}$ is the same-day open–close return, $\mathbb{1}_{i,t}^{h}$, $h \in \{\text{Small}, \text{Domestic}\}$, are indicator variables denoting whether a firm is of small capitalization or is domestically owned. Small and large stocks are defined as in Table 1. Low foreign ownership stocks are those outside the top quartile of ownership, or less than 30% ownership. Column (1) reports the regression for small-cap firms, and Column (2) for large-cap firms. Column (3) presents the interaction by market capitalization. Column (4) reports the regression for firms with high domestic ownership, and Column (5) for firms with high foreign ownership (foreign ownership share $\geq$ 80-percentile). Column (6) presents the interaction by ownership type. Column (7) includes both small-cap and low foreign ownership interactions. All specifications contains all relevant interacted indicator variables (e.g., $\mathbb{1}_{i,t}^{h} \times \mathbb{1}_{i,t}^{AE}$), which we omit from the table for brevity. Standard errors are two-way clustered by firm and date. ***, **, * corresponds to $p < .01$, $p < 0.05$, and $p < .1$, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$\mathbb{1}^{\text{AE}} \times \widehat{R}^{\text{AE, pre}}$	0.262***	-0.015	-0.038	0.275***	0.069	0.016	-0.063
	(0.08)	(0.05)	(0.06)	(0.07)	(0.06)	(0.06)	(0.06)
$\mathbb{1}^{\text{AJ}} \times \widehat{R}^{\text{AJ, pre}}$	0.225***	-0.023	-0.023	0.264***	0.051	0.013	-0.058
	(0.07)	(0.06)	(0.06)	(0.08)	(0.06)	(0.06)	(0.06)
$\mathbb{1}^{\text{UE}} \times \widehat{R}^{\text{UE, pre}}$	0.150*	0.072	0.038	0.108	0.123	0.078	0.028
	(0.08)	(0.07)	(0.07)	(0.08)	(0.07)	(0.08)	(0.07)
$\mathbb{1}^{\text{UJ}} \times \widehat{R}^{\text{UJ, pre}}$	-0.056	-0.266**	-0.474***	-0.198	-0.162	-0.369***	-0.506***
	(0.12)	(0.11)	(0.11)	(0.12)	(0.12)	(0.13)	(0.12)
$\mathbb{1}^{\text{Small}} \times \mathbb{1}^{\text{AE}} \times \widehat{R}^{\text{AE, pre}}$			0.355***				0.257**
			(0.09)				(0.12)
$\mathbb{1}^{\text{Small}} \times \mathbb{1}^{\text{AJ}} \times \widehat{R}^{\text{AJ, pre}}$			0.310***				0.199
			(0.10)				(0.13)
$\mathbb{1}^{\text{Small}} \times \mathbb{1}^{\text{UE}} \times \widehat{R}^{\text{UE, pre}}$			0.136				0.132
			(0.10)				(0.13)
$\mathbb{1}^{\text{Small}} \times \mathbb{1}^{\text{UJ}} \times \widehat{R}^{\text{UJ, pre}}$			0.427***				0.348*
			(0.15)				(0.20)
$\mathbb{1}^{\text{Domestic}} \times \mathbb{1}^{\text{AE}} \times \widehat{R}^{\text{AE, pre}}$						0.319***	0.145
						(0.09)	(0.12)
$\mathbb{1}^{\text{Domestic}} \times \mathbb{1}^{\text{AJ}} \times \widehat{R}^{\text{AJ, pre}}$						0.298***	0.168
						(0.09)	(0.12)
$\mathbb{1}^{\text{Domestic}} \times \mathbb{1}^{\text{UE}} \times \widehat{R}^{\text{UE, pre}}$						0.072	0.013
						(0.10)	(0.13)
$\mathbb{1}^{\text{Domestic}} \times \mathbb{1}^{\text{UJ}} \times \widehat{R}^{\text{UJ, pre}}$						0.283*	0.119
						(0.17)	(0.22)
Log(mcap)	0.132***	-0.010	0.081***	0.097***	0.004	0.052***	0.077***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Beta	-0.018	-0.081**	-0.030	-0.011	-0.053	-0.030	-0.031
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
N	118,389	54,570	172,959	109,337	63,622	172,959	172,959
T	2,534	2,620	2,624	2,483	2,623	2,624	2,624
R^2 (%)	0.39	0.05	0.38	0.36	0.03	0.35	0.38

Table 8: Readership Regressions for Japanese Stocks

This table reports the following regressions:

$$R_{i,t}^{OC} = a_t + b_0 \widehat{R}_{i,t}^{\text{EN}} + b_1 \widehat{R}_{i,t}^{\text{JP}} + b_2 \mathbb{1}_{i,t}^{\text{RE}} \times \widehat{R}_{i,t}^{\text{EN}} + b_3 \mathbb{1}_{i,t}^{\text{RJ}} \times \widehat{R}_{i,t}^{\text{JP}} + \text{Control}_{i,t} + \varepsilon_{i,t},$$

where $R_{i,t}^{O\to C}$ is the same-day open–close return, $\mathbb{1}_{i,t}^{\text{RE}}$ and $\mathbb{1}_{i,t}^{\text{RJ}}$ are indicator variables that equal one if, on day t, the English and Japanese news for stock i is read by at least 25 readers, conditional on the firm having news in that language. $\widehat{R}_{i,t}^{l}$ is the return score in language l. Columns (3)–(4) repeat the same experiment on aligned/unaligned news. Columns (2) and (4) add the control variables to present the full specification. All specifications contain all relevant interacted indicator variables (e.g., $\mathbb{1}_{i,t}^{\text{RE}} \times \mathbb{1}_{i,t}^{\text{AE}}$), which we omit from the table for brevity. Standard errors are two-way clustered by firm and date. The readership data begin in 2014. ***, **, * corresponds to $p < .01$, $p < 0.05$, and $p < .1$, respectively.

	(1)	(2)	(3)	(4)
$\widehat{R}^{\text{EN, pre}}$	0.339***	0.179***		
	(0.05)	(0.05)		
$\widehat{R}^{\text{JP, pre}}$	0.641***	0.543***		
	(0.05)	(0.05)		
$\mathbb{1}^{\text{AE}} \times \widehat{R}^{\text{AE, pre}}$			0.368***	0.265***
			(0.06)	(0.06)
$\mathbb{1}^{\text{AJ}} \times \widehat{R}^{\text{AJ, pre}}$			0.368***	0.275***
			(0.06)	(0.06)
$\mathbb{1}^{\text{UE}} \times \widehat{R}^{\text{UE, pre}}$			0.226***	0.131**
			(0.06)	(0.06)
$\mathbb{1}^{\text{UJ}} \times \widehat{R}^{\text{UJ, pre}}$			0.008	-0.031
			(0.11)	(0.10)
$\mathbb{1}^{\text{RE}} \times \widehat{R}^{\text{EN}}$	-0.391***	-0.204**		
	(0.09)	(0.09)		
$\mathbb{1}^{\text{RJ}} \times \widehat{R}^{\text{JP}}$	-0.463***	-0.401***		
	(0.08)	(0.08)		
$\mathbb{1}^{\text{RE}} \times \mathbb{1}^{\text{AE}} \times \widehat{R}^{\text{AE}}$			-0.435***	-0.343***
			(0.09)	(0.10)
$\mathbb{1}^{\text{RJ}} \times \mathbb{1}^{\text{AJ}} \times \widehat{R}^{\text{AJ}}$			-0.403***	-0.329***
			(0.10)	(0.10)
$\mathbb{1}^{\text{RE}} \times \mathbb{1}^{\text{UE}} \times \widehat{R}^{\text{UE}}$			-0.141	-0.045
			(0.13)	(0.12)
$\mathbb{1}^{\text{RJ}} \times \mathbb{1}^{\text{UJ}} \times \widehat{R}^{\text{UJ}}$			-0.546***	-0.577***
			(0.18)	(0.19)
Log(mcap)				0.054***
				(0.01)
Beta		-0.040**		-0.035
		(0.02)		(0.03)
N	340,309	331,290	177,995	172,959
T	2,627	2,627	2,624	2,624
R^2 (%)	0.29	0.36	0.30	0.36

Table 9: Alignment Statistics and Regression for US Stocks

Panel A of this table reports the return correlation as in Table 5. Panel B of this table reports the regression coefficients and errors of the specification reported in Table 6, but for US stocks. ***, **, * corresponds to $p < .01$, $p < 0.05$, and $p < .1$, respectively.

A. Return Score Correlation

	$\widehat{R}^{\text{AE}}$	$\widehat{R}^{\text{UE}}$	$\widehat{R}^{\text{EN}}$	$\widehat{R}^{\text{AJ}}$	$\widehat{R}^{\text{UJ}}$	$\widehat{R}^{\text{JP}}$
$\widehat{R}^{\text{AE}}$	1.00	0.20	0.58	0.64	0.21	0.58
$\widehat{R}^{\text{UE}}$	0.20	1.00	0.68	0.21	0.23	0.22
$\widehat{R}^{\text{EN}}$	0.58	0.68	1.00	0.45	0.20	0.45
$\widehat{R}^{\text{AJ}}$	0.64	0.21	0.45	1.00	0.22	0.79
$\widehat{R}^{\text{UJ}}$	0.21	0.23	0.20	0.22	1.00	0.44
$\widehat{R}^{\text{JP}}$	0.58	0.22	0.45	0.79	0.44	1.00

B. Return Forecasting

	(1)	(2)	(3)	(4)	(5)
$\mathbb{1}^{\text{AE}}$	-0.028				
	(0.04)				
$\mathbb{1}^{\text{AJ}}$		-0.029		-0.022	-0.023
		(0.04)		(0.04)	(0.04)
$\mathbb{1}^{\text{UE}}$	-0.009			-0.028	-0.156**
	(0.06)			(0.07)	(0.07)
$\mathbb{1}^{\text{UJ}}$		0.018		0.035	0.004
		(0.04)		(0.05)	(0.05)
$\mathbb{1}^{\text{AE}} \times \widehat{R}^{\text{AE, pre}}$	0.415***		0.423***	0.420***	0.384***
	(0.08)		(0.10)	(0.10)	(0.10)
$\mathbb{1}^{\text{AJ}} \times \widehat{R}^{\text{AJ, pre}}$		0.305***	-0.013	-0.013	0.004
		(0.08)	(0.10)	(0.10)	(0.09)
$\mathbb{1}^{\text{UE}} \times \widehat{R}^{\text{UE, pre}}$	0.115		0.112	0.109	0.078
	(0.17)		(0.16)	(0.16)	(0.16)
$\mathbb{1}^{\text{UJ}} \times \widehat{R}^{\text{UJ, pre}}$		0.170	0.060	0.089	-0.058
		(0.24)	(0.23)	(0.23)	(0.23)
Log(mcap)					0.051***
					(0.02)
Beta					-0.096
					(0.07)
N	51,321	51,321	51,321	51,321	49,586
T	2,703	2,703	2,703	2,703	2,703
R^2 (%)	0.07	0.03	0.07	0.08	0.20

Table 10: Robustness

Panels A and B report the regression results from Table 4 and Table 6 using data after June 1, 2020—the cutoff date for the LaBSE training data. Panel C uses the full sample starting in 2014 but restricts the analysis to one-to-one sentence alignments. Panels D-E of this table report the regression coefficients and errors of the specification reported in Table 4, with predicted returns constructed from text embeddings generated using GTE and E5. ***, **, * corresponds to $p < .01$, $p < 0.05$, and $p < .1$, respectively.

A. Post-2020

	(1)	(2)	(3)	(4)
$\widehat{R}^{\text{EN, pre}}$	0.761***		0.504***	0.301***
	(0.07)		(0.07)	(0.07)
$\widehat{R}^{\text{JP, pre}}$		0.731***	0.506***	0.486***
		(0.07)	(0.07)	(0.06)
Log(mcap)				0.026***
				(0.00)
Beta				-0.012
				(0.03)
N	142,818	142,818	142,818	139,111
T	1,064	1,064	1,064	1,064
R^2 (%)	0.16	0.17	0.23	0.27

B. Post-2020 Alignment

	(1)	(2)	(3)	(4)	(5)
$\mathbb{1}^{\text{AE}}$	-0.085*				
	(0.05)				
$\mathbb{1}^{\text{AJ}}$		-0.087*		-0.062	-0.006
		(0.05)		(0.05)	(0.05)
$\mathbb{1}^{\text{UE}}$	0.005			0.021	-0.013
	(0.06)			(0.06)	(0.06)
$\mathbb{1}^{\text{UJ}}$		0.029		0.022	-0.036
		(0.03)		(0.03)	(0.03)
$\mathbb{1}^{\text{AE}} \times \widehat{R}^{\text{AE, pre}}$	0.530***		0.330***	0.320***	0.221**
	(0.09)		(0.09)	(0.09)	(0.09)
$\mathbb{1}^{\text{AJ}} \times \widehat{R}^{\text{AJ, pre}}$		0.504***	0.278***	0.279***	0.229***
		(0.08)	(0.08)	(0.08)	(0.08)
$\mathbb{1}^{\text{UE}} \times \widehat{R}^{\text{UE, pre}}$	0.496***		0.316***	0.318***	0.182
	(0.11)		(0.11)	(0.11)	(0.12)
$\mathbb{1}^{\text{UJ}} \times \widehat{R}^{\text{UJ, pre}}$		0.554***	0.334**	0.362***	0.124
		(0.13)	(0.13)	(0.14)	(0.14)
Log(mcap)					0.040***
					(0.01)
Beta					-0.021
					(0.05)
N	88,165	88,165	88,165	88,165	85,658
T	1,061	1,061	1,061	1,061	1,061
R^2 (%)	0.16	0.15	0.18	0.18	0.24

C. Japan Mutual Top Alignment

	(1)	(2)	(3)	(4)	(5)
$\mathbb{1}^{\text{AE}}$	-0.024				
	(0.02)				
$\mathbb{1}^{\text{AJ}}$		-0.026		-0.001	0.055**
		(0.02)		(0.02)	(0.02)
$\mathbb{1}^{\text{UE}}$	0.010			0.008	-0.002
	(0.05)			(0.05)	(0.05)
$\mathbb{1}^{\text{UJ}}$		0.063**		0.041	-0.035
		(0.03)		(0.03)	(0.03)
$\mathbb{1}^{\text{AE}} \times \widehat{R}^{\text{AE, pre}}$	0.506***		0.319***	0.316***	0.198***
	(0.06)		(0.06)	(0.06)	(0.05)
$\mathbb{1}^{\text{AJ}} \times \widehat{R}^{\text{AJ, pre}}$		0.590***	0.319***	0.322***	0.220***
		(0.05)	(0.05)	(0.05)	(0.05)
$\mathbb{1}^{\text{UE}} \times \widehat{R}^{\text{UE, pre}}$	0.351***		0.271***	0.264***	0.148***
	(0.05)		(0.05)	(0.05)	(0.05)
$\mathbb{1}^{\text{UJ}} \times \widehat{R}^{\text{UJ, pre}}$		0.224***	0.066	0.096	-0.128*
		(0.07)	(0.07)	(0.07)	(0.07)
Log(mcap)					0.058***
					(0.01)
Beta					-0.030
					(0.03)
N	177,995	177,995	177,995	177,995	172,959
T	2,624	2,624	2,624	2,624	2,624
R^2 (%)	0.17	0.15	0.19	0.20	0.30

D. Japanese stocks (GTE/E5)

	GTE				E5			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\widehat{R}^{\text{EN, pre}}$	0.676***		0.433***	0.209***	0.685***		0.337***	0.118**
	(0.05)		(0.05)	(0.05)	(0.05)		(0.05)	(0.05)
$\widehat{R}^{\text{JP, pre}}$		0.741***	0.578***	0.440***		0.705***	0.580***	0.448***
		(0.05)	(0.06)	(0.06)		(0.05)	(0.05)	(0.05)
Log(mcap)				0.041***				0.041***
				(0.00)				(0.00)
Beta				-0.040**				-0.036*
				(0.02)				(0.02)
N	340,309	340,309	340,309	331,290	340,309	340,309	340,309	331,290
T	2,627	2,627	2,627	2,627	2,627	2,627	2,627	2,627
R^2 (%)	0.12	0.16	0.20	0.30	0.11	0.19	0.21	0.30

E. U.S. stocks (GTE/E5)

	GTE				E5			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\widehat{R}^{\text{EN, pre}}$	0.742***		0.735***	0.590***	0.748***		0.743***	0.597***
	(0.04)		(0.04)	(0.03)	(0.04)		(0.04)	(0.03)
$\widehat{R}^{\text{JP, pre}}$		0.598***	0.328***	0.443***		0.597***	0.276***	0.368***
		(0.08)	(0.08)	(0.08)		(0.08)	(0.08)	(0.08)
Log(mcap)				0.047***				0.045***
				(0.01)				(0.01)
Beta				-0.082***				-0.078***
				(0.02)				(0.02)
N	934,742	934,742	934,742	887,596	934,742	934,742	934,742	887,596
T	2,705	2,705	2,705	2,705	2,705	2,705	2,705	2,705
R^2 (%)	0.11	0.00	0.11	0.19	0.13	0.00	0.13	0.20

Table 11: Alignment Alphas

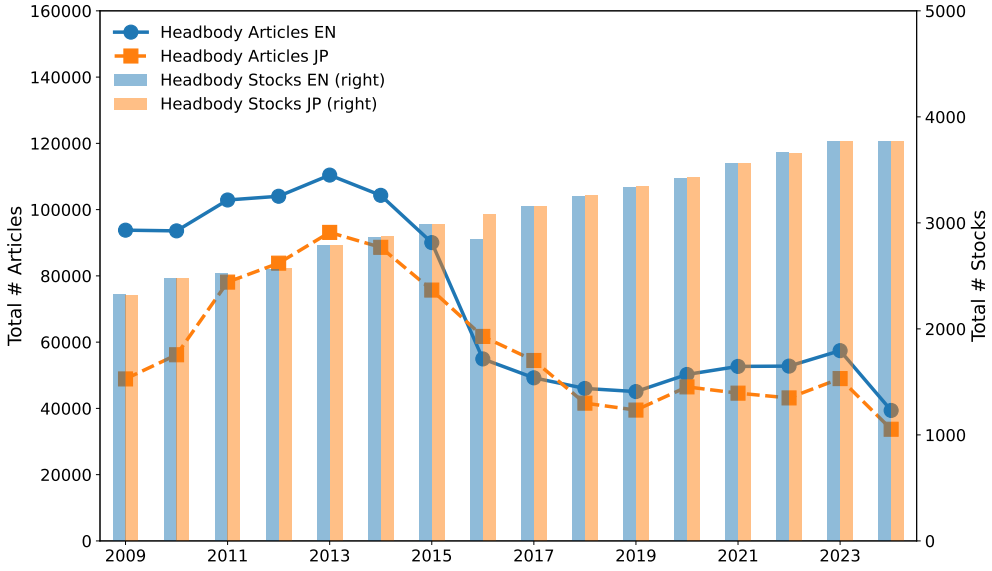
This table reports average returns and alphas of a double-sorted portfolio strategy analogous to Figure 8, where we sort stocks into quartiles using a composite aligned (AE+AJ) signal and UE signal. The strategy is long stocks that are in the highest quartiles and short stocks in the lowest quartiles of both signals. Returns and alphas are reported as daily percentage points. ***, **, * corresponds to $p < .01$, $p < 0.05$, and $p < .1$, respectively.

	Short	Long	Long-Short
Average Return	-0.515***	-0.024	0.491***
Standard Deviation	1.711	1.356	1.952
CAPM Alpha	-0.520***	-0.030	0.490***
FF5+mom Alpha	-0.518***	-0.029	0.489***

Figure 1: Summary Statistics

This figure shows the annual number of articles (left y-axis, articles with headline+body) and the number of stocks covered (right y-axis) from 2009 to 2024. Panel A reports for Japan, and Panel B reports for the U.S.

A. Japan



B. U.S.

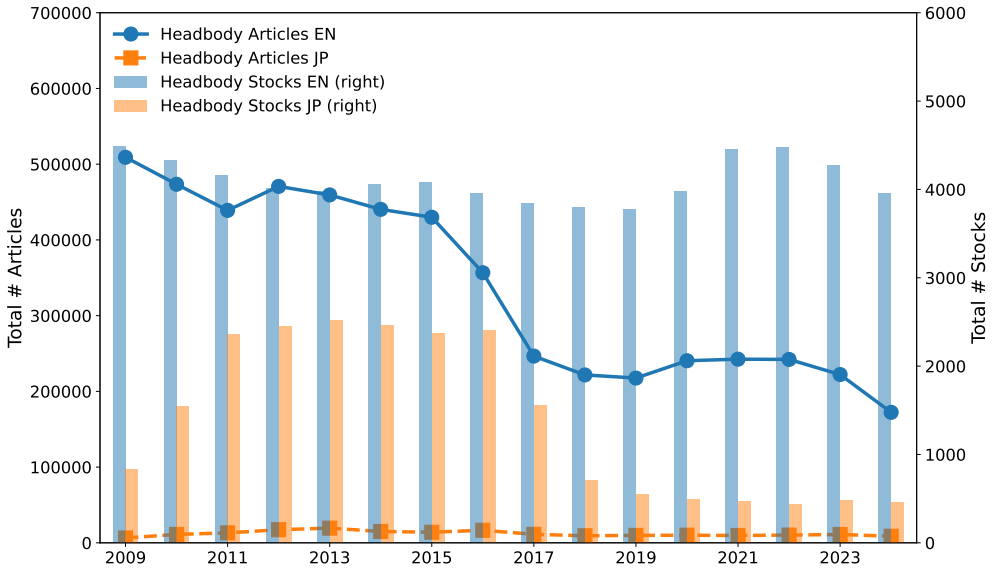
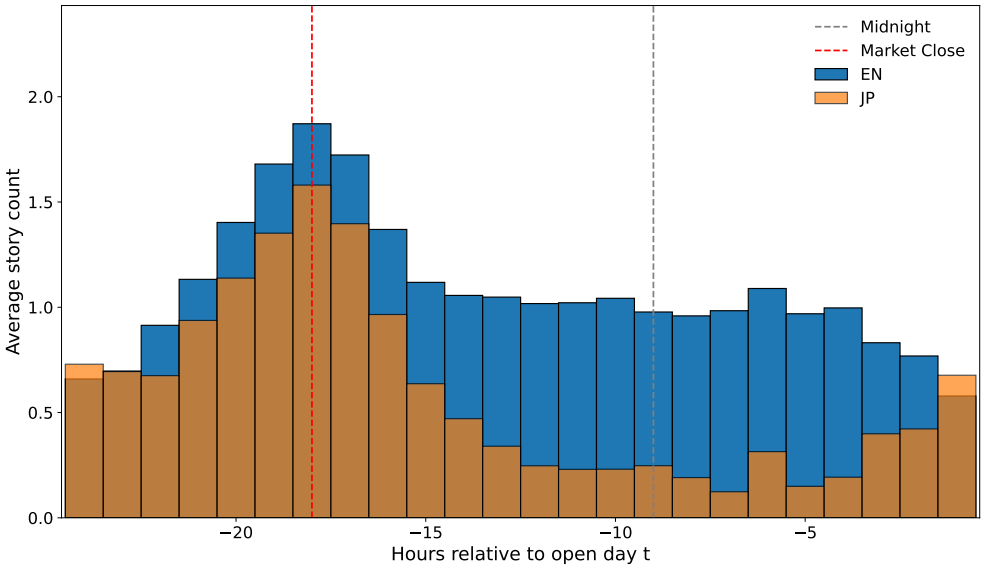


Figure 2: Production of Articles with Body Text

This figure plots the average number of news story bodies per stock released in each hour prior to the market open on day t. The cutoff is at the market open on day $t-1$. Blue bars represent English articles, and orange bars represent Japanese articles. Panel A shows results for Japan, while Panel B shows results for the U.S.

A. Japan



B. U.S.

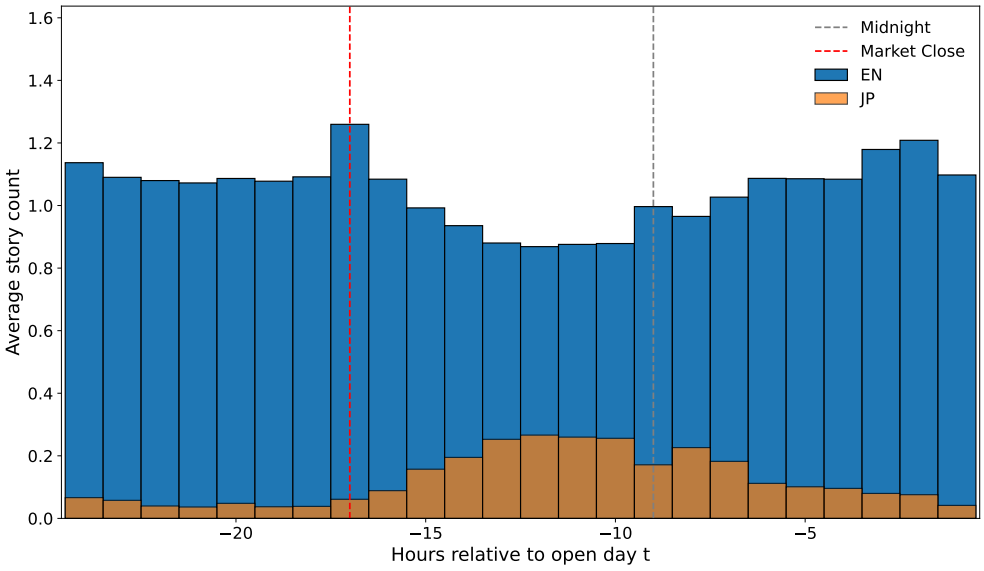


Figure 3: Timeline of Article Production

This figure shows the timing of news production relative to the trading day and the associated open–close, close–open, and open–open return windows.

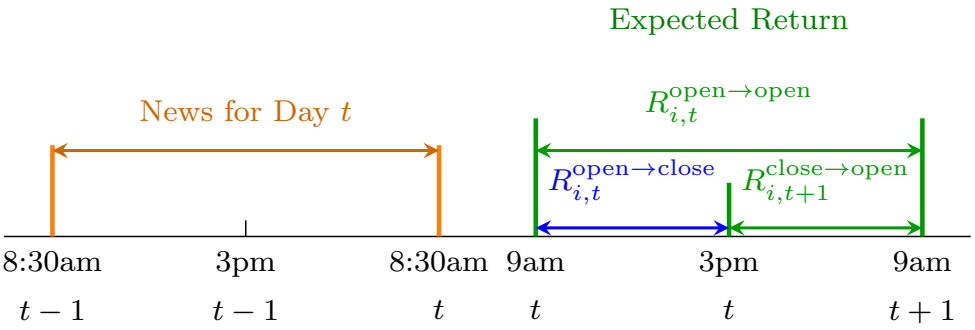


Figure 4: Alignment Examples

Panel A. shows the alignment between English and Japanese news for Mitsubishi UFJ Financial Group, Inc. (8306.T) on 2024-04-23. Panel B. shows the alignment between English and Japanese news for Apple (AAPL) on 2024-01-08. Translations of the Japanese articles are provided by Google Translate. Aligned sentences are underlined in pairs.

A. Mitsubishi UFJ Financial Group, Inc.

Type	Article
English	Mitsubishi UFJ Financial Group Inc. is considering sweetening its $1.7 billion offer for a minority stake in HDFC Bank Ltd.'s consumer lending unit in India,
	with a deal potentially hinging on valuation and the Japanese bank having a say in strategy, according to people familiar with the matter.
	The two sides are seeking to reach an agreement in about a month, the people said, asking not to be identified discussing private information.
	MUFG has been working on the purchase of a 20% stake in HDB Financial Services Ltd., Bloomberg News reported on April 12.
	A potential premium on such an investment is limited given MUFG would only be buying a minority stake in the business that offers personal, gold and auto loans, the people said.
	Still, MUFG sees the prospective deal as a way to form a long-term partnership with HDBF and build its presence in India, the people said.
	Final details under discussion include MUFG's influence on HDBF's strategy, the people said.
	Talks are ongoing and could face delays or not lead to a deal, the people said. Other bidders could also emerge, they said.
	A representative for MUFG declined to comment, while a spokesperson for HDFC didn't immediately respond to queries.
	HDFC's shares have dropped 11% this year, valuing the company at about $138 billion. MUFG is up 28% in that time and has a market value of $123 billion.
	HDFC, India's biggest lender, owns a 95% stake in HDBF and has been seeking ways to sell part of it, including a possible initial public offering.
	At an earnings briefing Saturday, HDFC Chief Financial Officer Srinivasan Vaidyanathan said HDBF needs to list by September 2025.
	He didn't comment on the MUFG talks, beyond saying any decision would be a financial one and all opportunities were being explored.
	India has become one of the busiest markets for deals, particularly as business in China dries up.
	MUFG has a branch in Gujarat International Finance Tec-City, a flagship project of Prime Minister Narendra Modi, along with a fund targeting middle- to late-stage Indian startups.
	It also has a stake in fintech company DMI Finance Pvt.
Japanese	三菱ＵＦＪフィナンシャル・グループ（ＭＵＦＧ）はインドの民間最大手ＨＤＦＣ銀行グループのノンバンクの少数持ち分取得に向け、条件の引き上げを検討している。事情に詳しい関係者が明らかにした。
	両社は約１カ月内に合意に達することを目指していると、関係者が非公開の情報だとして匿名を条件に述べた。
	ＭＵＦＧは 17 億ドル（約 2600 億円）程度を投じ、「ＨＤＢファイナンシャル・サービス（ＨＤＢＦＳ）」の株式 20%程度の取得に動いていると、ブルームバーグ・ニュースが 12 日に報じていた。
	取得が少数持ち分でしかないことから条件が引き上げられるとしても小幅だろうと、関係者は語った。
	ＨＤＢＦＳは消費者ローンや自動車ローンを提供している。
	ＭＵＦＧ、インド最大手銀ノンバンクに 2600 億円出資協議－関係者-関係者によると、ＭＵＦＧはＨＤＢＦＳと長期的なパートナーシップを築き、インドでの存在感を高める手段として、この取引を考えている。
	最終的な詰めの協議には、ＨＤＢＦＳの戦略に対するＭＵＦＧの影響力も含まれるという。
	関係者によれば、交渉は現在進行中で、合意が遅れる可能性や合意に至らない可能性もあり、他の買い手候補が現れる可能性もある。
	ＭＵＦＧの担当者はコメントを控え、ＨＤＦＣの広報担当者はコメント要請に直ちには返答しなかった。
Translated	Mitsubishi UFJ Financial Group (MUFG) is considering raising the terms of its bid to acquire a minority stake in a non-banking unit of India's largest private bank, HDFC Bank Group, according to people familiar with the matter.
	The two companies aim to reach an agreement within about a month, said the people, who asked not to be identified as the information is private.
	Bloomberg News reported on the 12th that MUFG is investing around $1.7 billion (approximately 260 billion yen) to acquire around 20% of HDB Financial Services (HDBFS).
	The acquisition is for a minority stake, so any increase in terms would likely be small, the people said.
	HDBFS offers consumer and auto loans.
	MUFG in talks to invest 260 billion yen in India's largest non-bank bank - sources said MUFG sees the deal as a way to build a long-term partnership with HDBFS and expand its presence in India.
	The final talks will also include MUFG's influence over HDBFS's strategy, the people said.
	Negotiations are ongoing and there is a possibility that an agreement could be delayed or not reached, or that other potential buyers could emerge, according to people familiar with the matter.
	A MUFG representative declined to comment, while a HDFC spokesman did not immediately respond to a request for comment.

Figure 4 continued.

B. Apple

Type	Article
English	Apple Inc.'s iPhone sales slump in China is deepening and the company is likely to see volumes decline further this year, according to Jefferies analysts led by Edison Lee. The iPhone maker's latest generation got off to an atypically sluggish start in China last year, which most recently expanded to a 30% year-on-year decline, Lee and colleagues said in a note on Sunday, citing industry checks. The rest of the country's mobile market grew in December, with Huawei Technologies Co. growing fastest on the back of its new Mate 60 device lineup. Weeks before the iPhone 15 went on sale in September, Huawei's debut of the Mate 60 Pro —which runs on a new made-in-China system processor —spurred a patriotic fervor that reclaimed some of the customers it previously lost to Apple. Jefferies estimates Huawei shipped 35 million smartphones in 2023, with some supply constraints preventing that number from being larger. Apple saw a double-digit fall in volumes in December and Jefferies forecasts a similar decline for 2024. Discounts on Apple's smartphone range increased last week across various online shopping portals, cutting into the average selling price without stimulating growth in volume. Apple gained share in China after US sanctions cut Huawei off from the world's leading chipmakers, such as Taiwan Semiconductor Manufacturing Co., in 2020. The Shenzhen-based electronics firm's return to competitiveness in the mobile market has seen it claw back market share, and it's now developing its own software ecosystem to compete with Apple's iOS and Alphabet Inc.'s Android.
Japanese	米アップルのスマートフォン「iPhone」の中国における販売低迷は深刻化しつつあり、今年さらに販売台数が減少する可能性が高い。 エディソン・リー氏率いるジェフリーズのアナリストがそうした見方を示した。 リー氏らは7日付のリポートで業界調査を引用し、最新モデル「iPhone 15」の低調な売れ行きが響き、最近の iPhone 販売は前年比で 30%減少していると分析。 一方、競合する中国の華為技術（ファーウェイ）は、「Mate 60 Pro」シリーズの好調で急速な伸びを示している。 ジェフリーズは中国における 2024 年の iPhone 販売について 2 桁台の減少を予想。 さまざまなオンラインショッピングポータルでアップルのスマートフォン製品の値引きが先週相次ぎ、平均販売価格が低下したものの、販売台数の伸びにはつながっていない。
Translated	Sales slumps for Apple's iPhone smartphones in China are becoming more severe, and unit sales are likely to fall further this year. This view was expressed by analysts at Jefferies led by Edison Li. In a report dated the 7th, Li et al. cited industry research, stating that recent iPhone sales have fallen 30% year-on-year due to sluggish sales of the latest model, the iPhone 15. Meanwhile, Chinese rival Huawei Technologies is experiencing rapid growth thanks to the success of its Mate 60 Pro series. Jefferies predicts a double-digit decline in iPhone sales in China in 2024. A series of discounts on Apple's smartphone products on various online shopping portals last week lowered average selling prices, but this has not translated into increased unit sales.

Figure 5: Alignment Statistics

Panel A reports the average sentence-level alignment proportions each year for Japan. For each trading day, we compute the proportion of aligned Japanese and aligned English sentences in each market, then average these measures of the year. Panel B reports the same averages for the U.S. market.

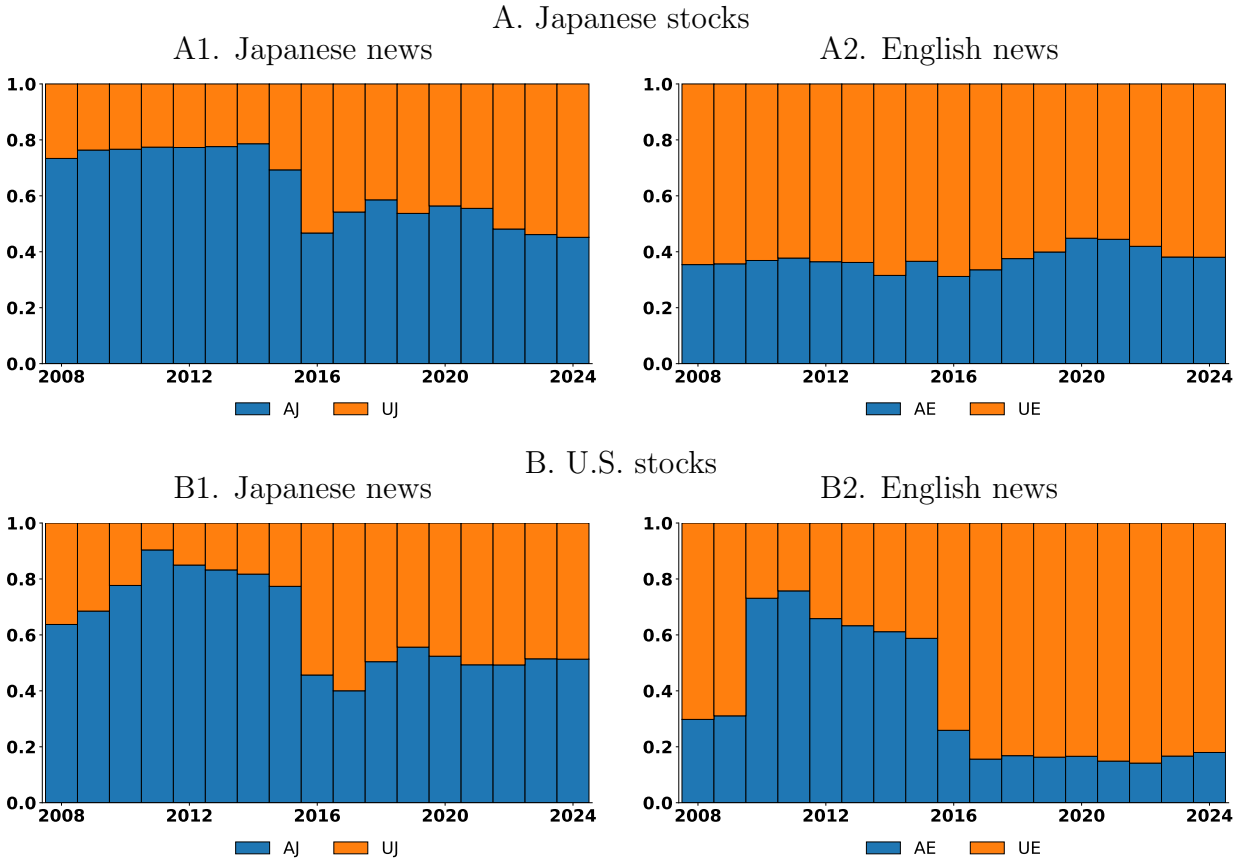
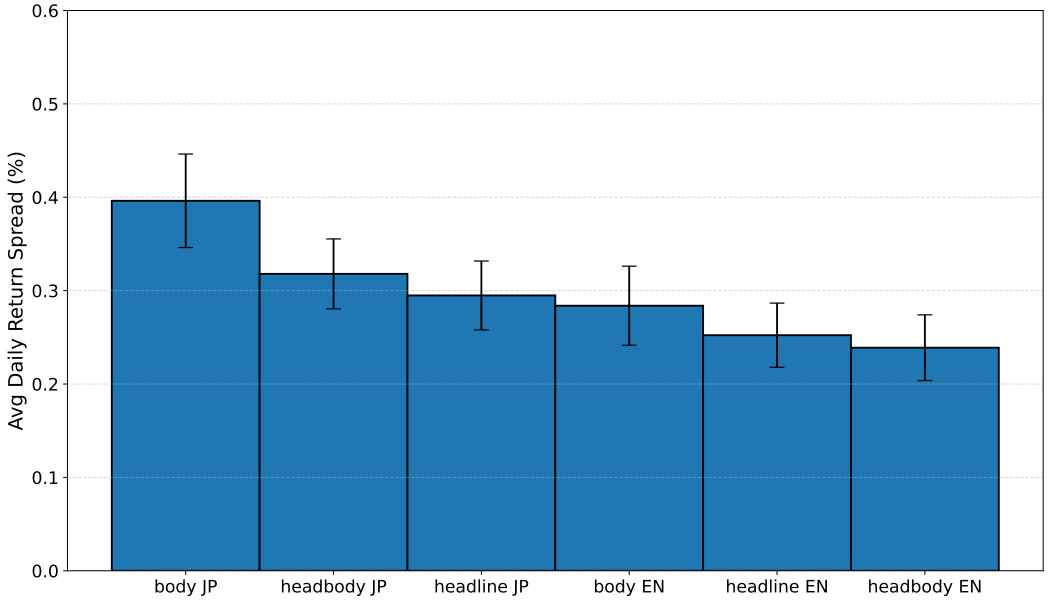


Figure 6: Model Performance

Panel A shows the average return spread between the long leg (top quantile) and short leg (bottom quantile) of $R_{i,t}^{O\to C}$, where the Japanese stocks are sorted by predicted returns $\widehat{R}_{i,t}^{\text{l, pre}}$ using LaBSE embeddings. Panel B reports the average return spreads for multiple horizons: same-day open–close $R_{i,t}^{O\to C} = \frac{\text{Close}_t - \text{Open}_t}{\text{Open}_t}$, close-to-next-open $R_{i,t}^{C\to O} = \frac{\text{Open}_{t+1} - \text{Close}_t}{\text{Close}_t}$, next-day open-to-open $R_{i,t+1}^{O\to O} = \frac{\text{Open}_{t+1} - \text{Open}_t}{\text{Open}_t}$, and $R_{i,t+2}^{O\to O} = \frac{\text{Open}_{t+2} - \text{Open}_{t+1}}{\text{Open}_{t+1}}$. $R_{i,t}^{C\to O}$, $R_{i,t+1}^{O\to O}$ and $R_{i,t+2}^{O\to O}$ are adjusted for dividends and stock splits. The returns are reported in percentage.

A. Average return spread comparison



B. Return spread decay

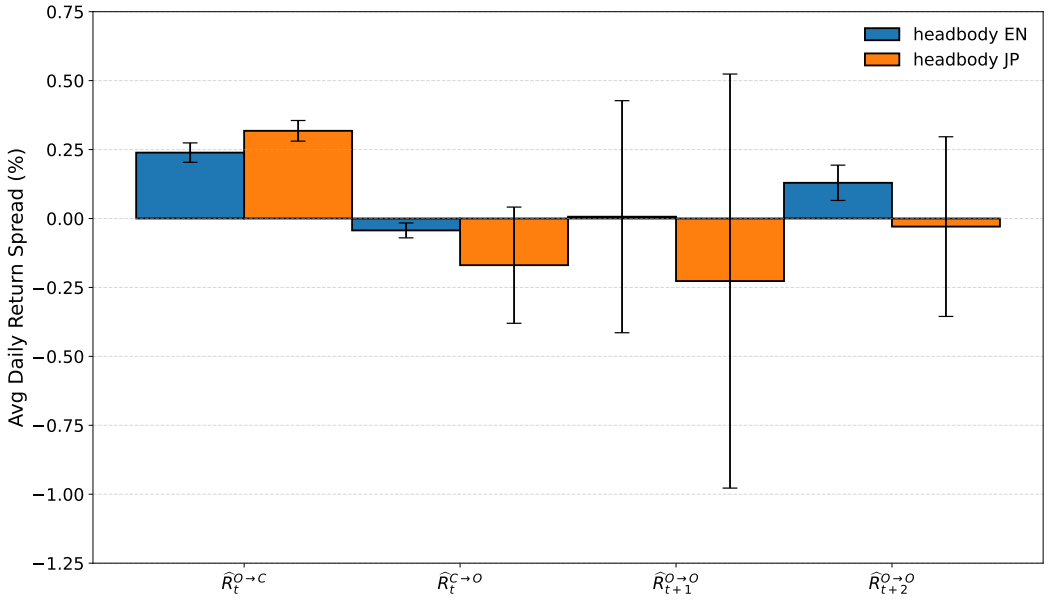
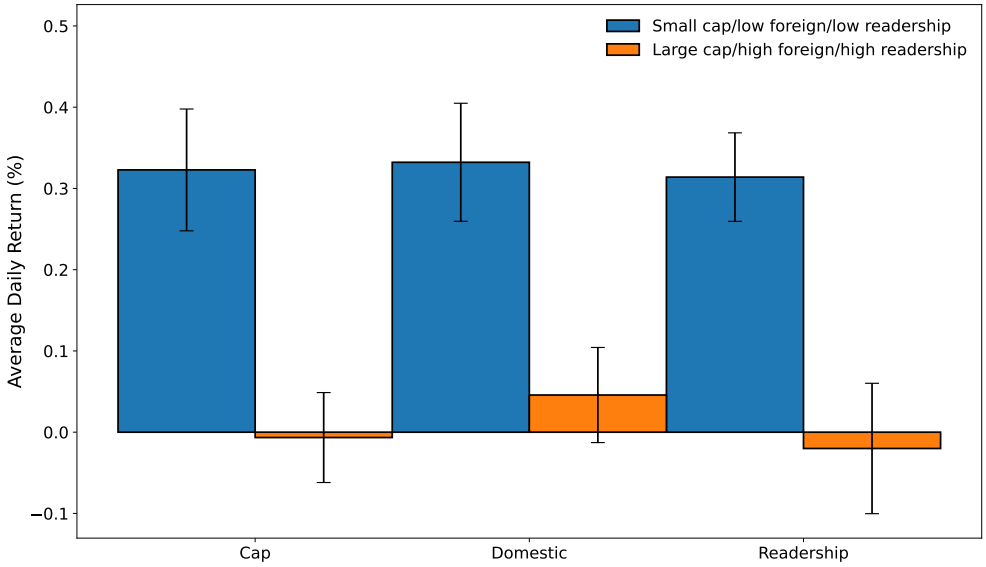


Figure 7: Reaction to aligned news

This figure reports the average daily long-short portfolio return to a strategy that buys Japanese stocks with $\widehat{R}_{i,t}^{\text{AJ, pre}}$ above the daily median and sells Japanese stocks below the median implemented within sub-samples of stocks, where we apply a common cutoff for each sub-sample. Small/large stocks, low/high foreign ownership, and low/high readership are defined in Tables 1, 7, and 8. Panel B repeats the exercise for aligned English scores $\widehat{R}_{i,t}^{\text{AE, pre}}$. Returns are reported in percentage points.

A. Aligned Japanese



B. Aligned English

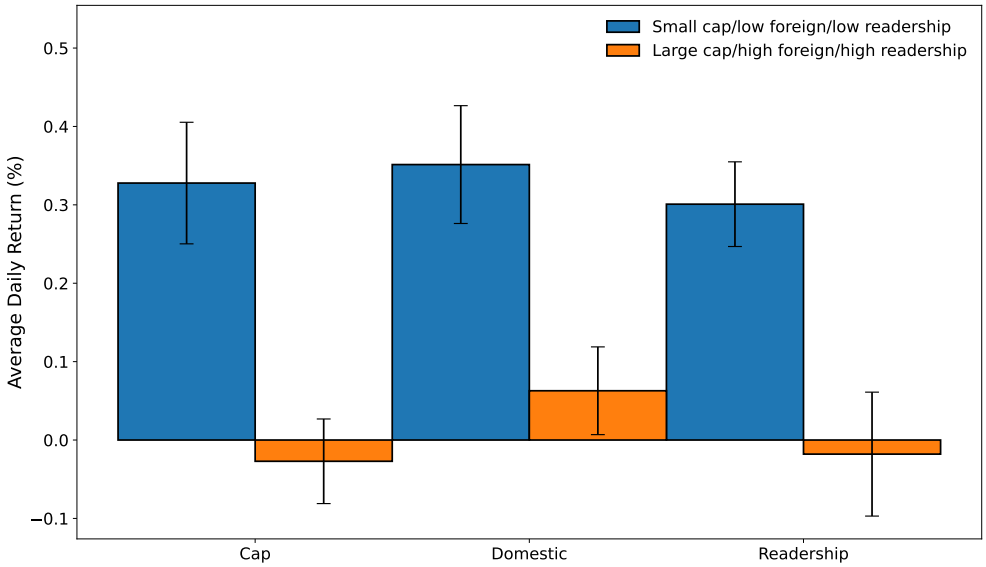
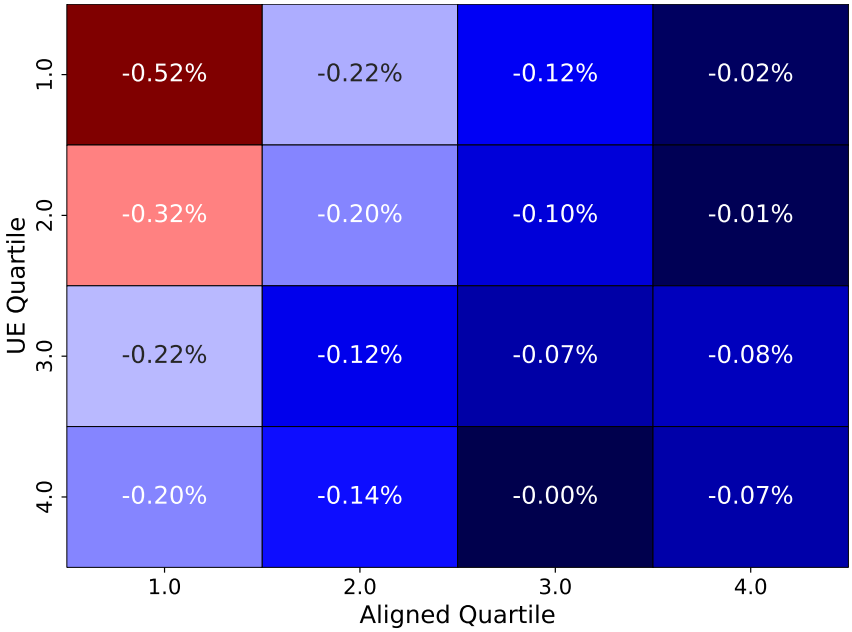


Figure 8: Alignment Return Double Sort

We compute the average alignment predicted return scores $\widehat{R}^{\text{A, pre}} = \frac{\widehat{R}^{\text{AE, pre}} + \widehat{R}^{\text{AJ, pre}}}{2}$. Panel A shows the average return heatmap, sorted by daily quartiles of $\widehat{R}^{\text{A, pre}}$ and $\widehat{R}^{\text{UE, pre}}$.



Measuring Price Effects of Multilingual Global News with Large Language Models

Internet Appendix

A Additional Tables & Figures

Table A.1: Article Length

This table reports the average length of each headline and body article for Japan and US. The length of Japanese articles are approximated by the number of Japanese characters, while the length of English articles are approximated by the number of English words (split by space).

Year	Japanese market news				U.S. market news			
	Headline		Body		Headline		Body	
	JP	EN	JP	EN	JP	EN	JP	EN
2008	16	9	121	111	20	9	288	231
2009	18	9	113	99	20	9	202	201
2010	17	9	116	92	14	9	91	185
2011	17	9	142	86	11	9	92	186
2012	16	9	144	92	14	9	161	197
2013	16	9	156	103	15	9	174	208
2014	16	9	162	97	15	9	172	192
2015	16	9	167	97	15	9	180	169
2016	16	9	204	157	16	9	191	169
2017	16	9	204	129	20	9	245	162
2018	16	9	212	118	22	9	332	171
2019	17	9	215	140	21	9	286	186
2020	17	8	206	130	21	9	312	194
2021	17	8	220	139	21	9	365	208
2022	17	8	233	157	20	9	403	293
2023	17	8	241	146	20	8	423	217
2024	17	8	243	147	20	8	443	218
Average	17	9	182	120	18	9	257	199

Table A.2: U.S. Alphas

This table reports the alphas for US, following Table 3.

A. CAPM Alpha (Headbody Japanese News)

	LaBSE			GTE			E5		
	Short	Long	Long-Short	Short	Long	Long-Short	Short	Long	Long-Short
Average Return	-0.198***	0.065*	0.263***	-0.199***	0.087**	0.286***	-0.208***	0.062*	0.270***
Standard Deviation	1.735	1.393	1.765	1.800	1.420	1.785	1.748	1.320	1.722
CAPM Alpha	-0.239***	0.026	0.265***	-0.242***	0.047	0.289***	-0.249***	0.021	0.270***
FF5+mom Alpha	-0.227***	0.033	0.261***	-0.230***	0.056*	0.285***	-0.239***	0.030	0.269***

B. CAPM Alpha (Headbody English News)

	LaBSE			GTE			E5		
	Short	Long	Long-Short	Short	Long	Long-Short	Short	Long	Long-Short
Average Return	-0.283***	0.049**	0.332***	-0.323***	0.069***	0.392***	-0.334***	0.076***	0.410***
Standard Deviation	1.417	1.082	0.994	1.438	1.071	1.038	1.480	1.032	1.042
CAPM Alpha	-0.321***	0.012	0.333***	-0.361***	0.033**	0.394***	-0.373***	0.040***	0.413***
FF5+mom Alpha	-0.297***	0.023*	0.320***	-0.336***	0.044***	0.380***	-0.347***	0.051***	0.398***

Table A.3: Sharpe Ratios by Model, Region, and Strategy

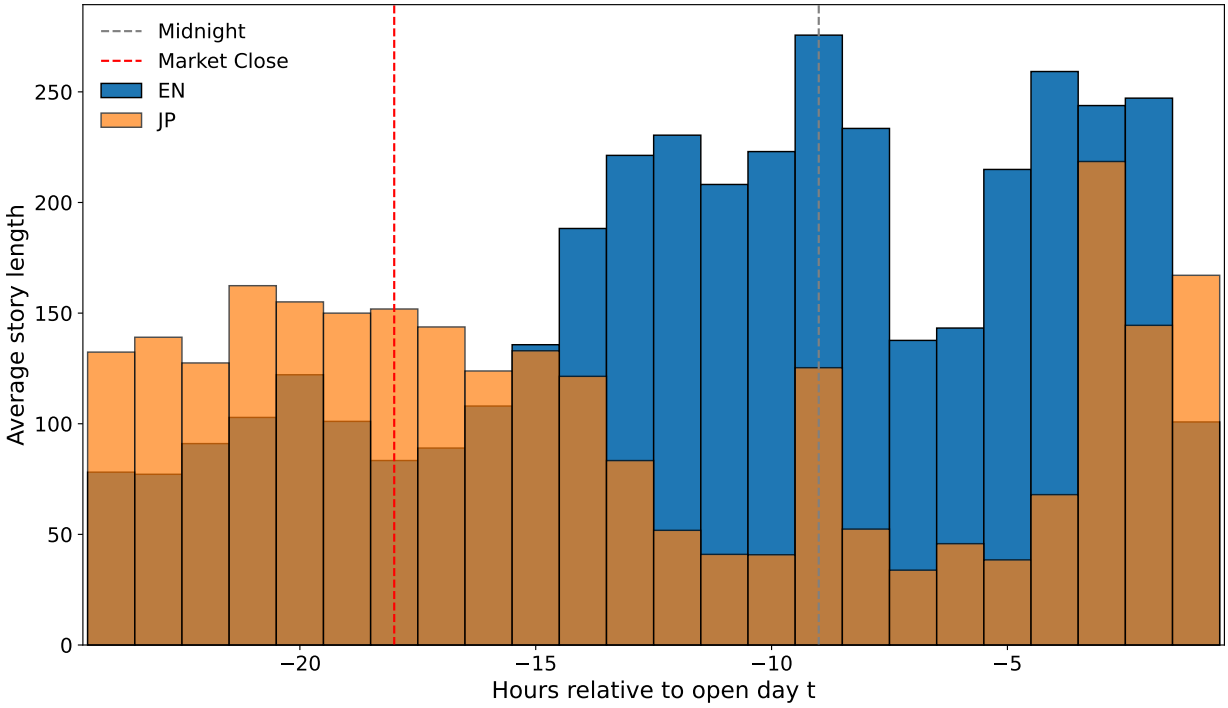
This table reports Sharpe ratios for best-performing strategies that account for transaction costs. We assume 10 basis points on each leg of a trade for large stocks and 20 basis points for small stocks, following Chen, Kelly, and Xiu (2022). For reference, the table also reports the best-performing strategies from Table 3 without transaction costs.

	JP		US	
	T-Costs	Table 3	T-Costs	Table 3
LaBSE	3.00	5.29	2.87	5.34
	JP (B)	JP (HB)	EN (HB)	EN (HL)
GTE	2.20	4.69	3.67	6.00
	EN (B)	EN (HB)	EN (HB)	EN (HB)
E5	2.51	5.03	3.93	6.26
	JP (HL)	JP (HL)	EN (HB)	EN (HL)

Figure A.1: Timeline of Article Production (Body Length)

This figure complements Figure 3 by plotting the average length of news stories released in each hour prior to the market open on day t. The cutoff is at the market open on day $t-1$. Panel A shows body-text lengths for the Japan market, and Panel B shows the corresponding results for the US market.

A. Japan



B. U.S.

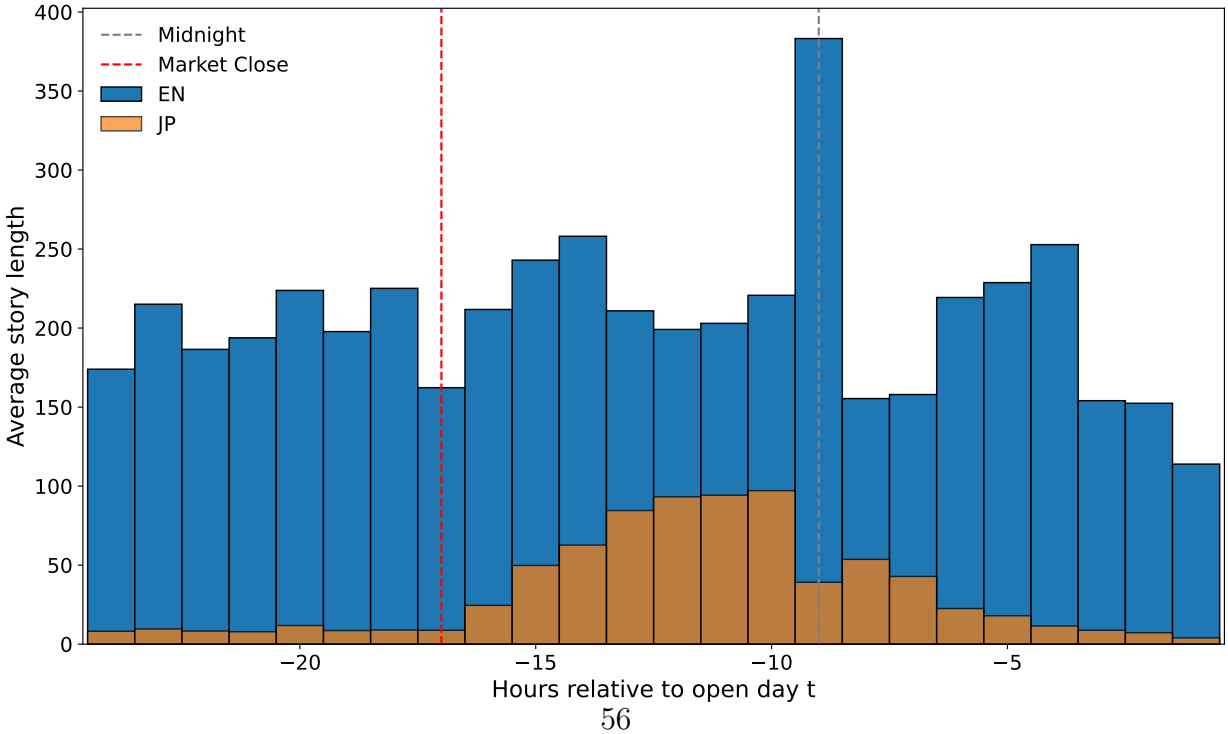
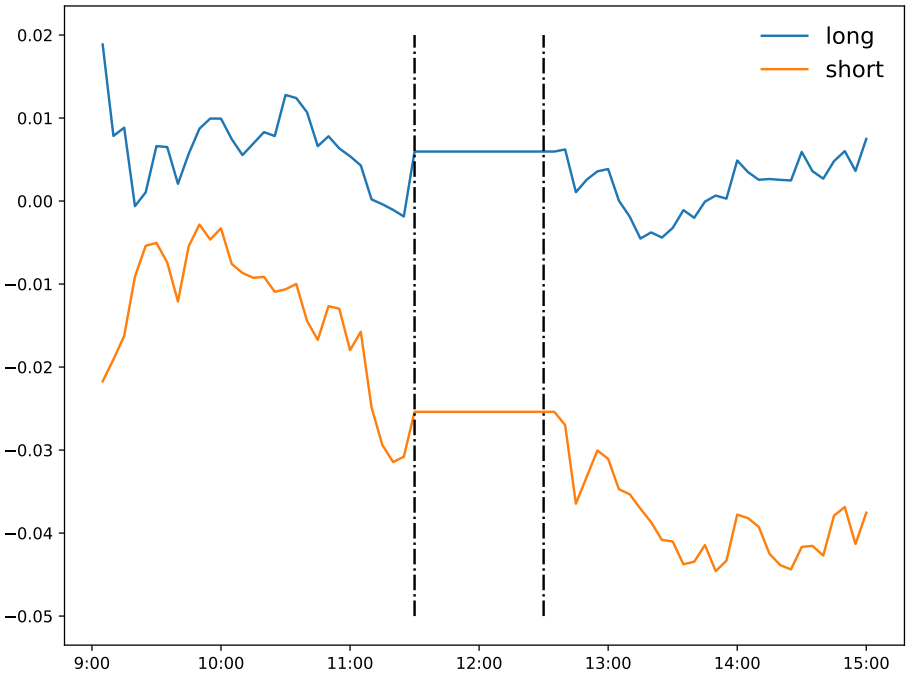


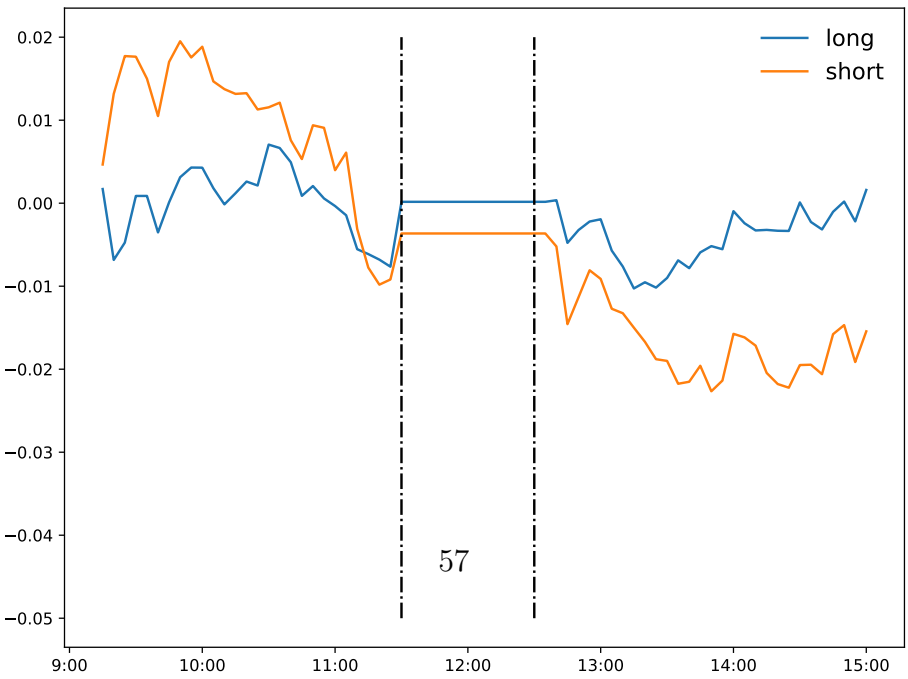
Figure A.2: Intraday Returns

From a subsample of 298 Japanese stocks with intraday ask and bid prices, we compute 5-minute returns using midquotes and cumulate them within each trading day. In this subsample, most of the stocks are the highly liquid with top market capitalization. Stocks are sorted daily based on the predicted open–close return score computed on Japanese headbody text ($\hat{R}^{\mathrm{EN,\ pre}}$) with embeddings from LaBSE. We form a long–short portfolio by going long (short) stocks with predicted scores in the top (bottom) quintiles. For each day, we compute cumulative intraday returns separately for the long and short legs, and then average these cumulative return paths across all days. During the midday trading break (11:30–12:35), returns are set to zero. On the y-axis, returns are expressed in percentage terms.

A. trade starting 9:05 (first return 9:00-9:05)



B. trade starting 9:15 (first return 9:10-9:15)



B Conceptual Framework

This Internet Appendix provides a formal version of the conceptual framework in Section 3. We (i) specify a simple data–generating process (DGP) for returns and multilingual news, (ii) derive the full–information benchmark, (iii) introduce language–segmented investors with variance–inflation factors, (iv) solve for equilibrium prices and return–prediction slopes, and (v) prove Predictions 1 and 2.

B.1 Environment and Data–Generating Process

We focus on the Japanese market; the U.S. case is analogous. There is a single risky asset with payoff x at date 1 and a risk–free asset normalized to zero return. Time 0 is "today", when news arrives and the asset is traded.

Factors and payoff. The terminal payoff is driven by three orthogonal news components and residual noise:

$$x = A + U_J + U_E + \varepsilon_x, \tag{1}$$

where

$$\begin{aligned} A &\sim \mathcal{N}(0, \sigma_A^2) && \text{common aligned news,} \\ U_J &\sim \mathcal{N}(0, \sigma_{U_J}^2) && \text{residual Japanese news,} \\ U_E &\sim \mathcal{N}(0, \sigma_{U_E}^2) && \text{residual English news,} \\ \varepsilon_x &\sim \mathcal{N}(0, \sigma_x^2) && \text{payoff noise,} \end{aligned}$$

and all shocks are mutually independent. Thus $\mathrm{Var}(x) = \sigma_A^2 + \sigma_{U_J}^2 + \sigma_{U_E}^2 + \sigma_x^2$.

Signals from news. Investors observe four signals summarizing Bloomberg coverage:

$$s \equiv \begin{pmatrix} s_{AJ} \\ s_{AE} \\ s_{UJ} \\ s_{UE} \end{pmatrix},$$

where A denotes aligned (common) content and U denotes unaligned (residual) content, with subscripts J and E indicating Japanese and English, respectively. We model each score as the sum of an underlying factor and measurement noise:

$$s_{AJ} = A + \varepsilon_{AJ}, \qquad s_{AE} = A + \varepsilon_{AE}, \tag{2}$$

$$s_{UJ} = U_J + \varepsilon_{UJ}, \qquad s_{UE} = U_E + \varepsilon_{UE}, \tag{3}$$

with

$$\varepsilon_{AJ} \sim \mathcal{N}(0, \sigma^2_{AJ}), \qquad \varepsilon_{AE} \sim \mathcal{N}(0, \sigma^2_{AE}),$$
$$\varepsilon_{UJ} \sim \mathcal{N}(0, \sigma^2_{e,U_J}), \qquad \varepsilon_{UE} \sim \mathcal{N}(0, \sigma^2_{e,U_E}).$$

All ε terms are independent of $(A, U_J, U_E, \varepsilon_x)$ and of each other.

Equations (1)–(3) imply that (x, s) is jointly Gaussian. Let $\Sigma \equiv \mathrm{Var}(s)$ and $\lambda \equiv \mathrm{Cov}(s, x)$. From the factor structure we obtain:

$$\lambda = \mathrm{Cov}(s, x) = \begin{pmatrix} \sigma^2_A \\ \sigma^2_A \\ \sigma^2_{U_J} \\ \sigma^2_{U_E} \end{pmatrix}, \tag{4}$$

and the covariance matrix of s has a convenient block–diagonal form,

$$\Sigma = \mathrm{Var}(s) = \begin{bmatrix} \Sigma_A & 0 \\ 0 & \Sigma_U \end{bmatrix}, \tag{5}$$

where the aligned block is

$$\Sigma_A = \mathrm{Var}\begin{pmatrix} s_{AJ} \\ s_{AE} \end{pmatrix} = \begin{bmatrix} \sigma^2_A + \sigma^2_{AJ} & \sigma^2_A \\ \sigma^2_A & \sigma^2_A + \sigma^2_{AE} \end{bmatrix}, \tag{6}$$

and the unaligned block is

$$\Sigma_U = \mathrm{Var}\begin{pmatrix} s_{UJ} \\ s_{UE} \end{pmatrix} = \begin{bmatrix} \sigma^2_{U_J} + \sigma^2_{e,U_J} & 0 \\ 0 & \sigma^2_{U_E} + \sigma^2_{e,U_E} \end{bmatrix}. \tag{7}$$

By construction, aligned and unaligned scores are independent conditional on (A, U_J, U_E), and unaligned scores are independent across languages.

B.2 Full–Information Omnilingual Benchmark

A hypothetical omnilingual Bayesian investor knows the true joint distribution of (x, s). In the linear–Gaussian setup, the posterior mean of x given s is the linear projection

$$\widehat{x}^{\text{full}}(s) = \mathrm{E}[x \mid s] = \omega^\top s, \qquad \omega \equiv \Sigma^{-1}\lambda. \tag{8}$$

The vector $\omega = (\omega_{AJ}, \omega_{AE}, \omega_{UJ}, \omega_{UE})^\top$ contains the full–information weights on each score.

Because Σ is block–diagonal, ω can be computed block–wise. For the aligned block, combining (6) with the standard 2×2 inverse formula yields

$$\Sigma_A^{-1} = \frac{1}{D_A}\begin{bmatrix} \sigma^2_A + \sigma^2_{AE} & -\sigma^2_A \\ -\sigma^2_A & \sigma^2_A + \sigma^2_{AJ} \end{bmatrix}, \qquad D_A \equiv \sigma^2_A(\sigma^2_{AJ} + \sigma^2_{AE}) + \sigma^2_{AJ}\sigma^2_{AE}. \tag{9}$$

Using $\lambda_A \equiv (\sigma_A^2, \sigma_A^2)^\top$ (the first two elements of λ in (4)) we obtain

$$\omega_{AJ} = e_1^\top \Sigma_A^{-1} \lambda_A = \frac{\sigma_A^2 \sigma_{AE}^2}{D_A}, \tag{10}$$

$$\omega_{AE} = e_2^\top \Sigma_A^{-1} \lambda_A = \frac{\sigma_A^2 \sigma_{AJ}^2}{D_A}, \tag{11}$$

where e_1, e_2 are the first two standard basis vectors. Both weights are strictly positive. Their sum is

$$\omega_{AJ} + \omega_{AE} = \frac{\sigma_A^2(\sigma_{AJ}^2 + \sigma_{AE}^2)}{D_A} = \frac{\sigma_A^2(\sigma_{AJ}^2 + \sigma_{AE}^2)}{\sigma_A^2(\sigma_{AJ}^2 + \sigma_{AE}^2) + \sigma_{AJ}^2 \sigma_{AE}^2} \in (0, 1). \tag{12}$$

Thus the omnilingual investor puts a total weight strictly between 0 and 1 on the aligned block (s_{AJ}, s_{AE}).

For the unaligned block, Σ_U in (7) is diagonal, so

$$\Sigma_U^{-1} = \begin{bmatrix} \frac{1}{\sigma_{U_J}^2 + \sigma_{e,U_J}^2} & 0 \\ 0 & \frac{1}{\sigma_{U_E}^2 + \sigma_{e,U_E}^2} \end{bmatrix}. \tag{13}$$

Let $\lambda_U \equiv (\sigma_{U_J}^2, \sigma_{U_E}^2)^\top$ collect the last two elements of (4). Then

$$\omega_{UJ} = \frac{\sigma_{U_J}^2}{\sigma_{U_J}^2 + \sigma_{e,U_J}^2}, \qquad \omega_{UE} = \frac{\sigma_{U_E}^2}{\sigma_{U_E}^2 + \sigma_{e,U_E}^2}. \tag{14}$$

Again, these are strictly between 0 and 1.

We emphasize that ω depends only on the objective covariance structure (Σ, λ); it does not depend on investor composition or language frictions.

B.3 Language–Segmented Investors

We now introduce two representative investor types:

- Local investors J (Japanese), who natively read Japanese news;
- Foreign (English–language) investors E, who natively read English news.

Let q_J, q_E denote their population shares with $q_J + q_E = 1$ and $q_J, q_E \in (\underline{q}, 1 - \underline{q})$ for some $\underline{q} > 0$.

Perceived covariance matrices. Both types know the basic factor structure but mis-perceive non–native language scores as noisier than they really are. We capture this using variance adjustment factors $\kappa_l^i > 0$ applied by investor i to language l. We keep the general notation throughout. In the baseline case, we assume $\kappa_E^J, \kappa_J^E > 1$ and $\kappa_J^J, \kappa_E^E = 1$ so that:

- Japanese investors J inflate the variances of English scores (s_{AE}, s_{UE});
- English investors E inflate the variances of Japanese scores (s_{AJ}, s_{UJ}).

Formally, their perceived covariance matrices are

$$\Sigma_J = \begin{bmatrix} \Sigma_A^{(J)} & 0 \\ 0 & \Sigma_U^{(J)} \end{bmatrix}, \qquad \Sigma_E = \begin{bmatrix} \Sigma_A^{(E)} & 0 \\ 0 & \Sigma_U^{(E)} \end{bmatrix}, \tag{15}$$

where the aligned blocks are

$$\Sigma_A^{(J)} = \begin{bmatrix} \sigma_A^2 + \kappa_J^J \sigma_{AJ}^2 & \sigma_A^2 \\ \sigma_A^2 & \sigma_A^2 + \kappa_E^J \sigma_{AE}^2 \end{bmatrix}, \tag{16}$$

$$\Sigma_A^{(E)} = \begin{bmatrix} \sigma_A^2 + \kappa_J^E \sigma_{AJ}^2 & \sigma_A^2 \\ \sigma_A^2 & \sigma_A^2 + \kappa_E^E \sigma_{AE}^2 \end{bmatrix}, \tag{17}$$

and the unaligned blocks are

$$\Sigma_U^{(J)} = \begin{bmatrix} \sigma_{U_J}^2 + \kappa_J^J \sigma_{e,U_J}^2 & 0 \\ 0 & \sigma_{U_E}^2 + \kappa_E^J \sigma_{e,U_E}^2 \end{bmatrix}, \tag{18}$$

$$\Sigma_U^{(E)} = \begin{bmatrix} \sigma_{U_J}^2 + \kappa_J^E \sigma_{e,U_J}^2 & 0 \\ 0 & \sigma_{U_E}^2 + \kappa_E^E \sigma_{e,U_E}^2 \end{bmatrix}. \tag{19}$$

In the baseline case, each type believes that foreign–language scores contain extra idiosyncratic noise independent of the underlying factors. Crucially, both types agree on $\lambda = \mathrm{Cov}(s, x)$ in (4), so their perceived joint distributions are internally consistent: adding independent noise raises variances but leaves covariances with x unchanged.

Posterior weights under segmentation. Investor $i \in \{J, E\}$ uses their perceived covariance matrix Σ_i to form a posterior mean for x given s,

$$\widehat{x}_i(s) = \mathrm{E}_i[x \mid s] = \theta_i^\top s, \qquad \theta_i \equiv \Sigma_i^{-1} \lambda. \tag{20}$$

The vector $\theta_i = (\theta_{AJ}^i, \theta_{AE}^i, \theta_{UJ}^i, \theta_{UE}^i)^\top$ collects the subjective GLS weights used by type i.

Because Σ_i is block–diagonal, θ_i again decomposes into aligned and unaligned parts. For Japanese investors, the aligned block is governed by $\Sigma_A^{(J)}$ in (16). Its inverse is

$$\left(\Sigma_A^{(J)}\right)^{-1} = \frac{1}{D_J} \begin{bmatrix} \sigma_A^2 + \kappa_E^J \sigma_{AE}^2 & -\sigma_A^2 \\ -\sigma_A^2 & \sigma_A^2 + \kappa_J^J \sigma_{AJ}^2 \end{bmatrix}, \tag{21}$$

with

$$D_J \equiv \sigma_A^2(\kappa_J^J \sigma_{AJ}^2 + \kappa_E^J \sigma_{AE}^2) + \kappa_J^J \kappa_E^J \sigma_{AJ}^2 \sigma_{AE}^2. \tag{22}$$

Multiplying by $\lambda_A = (\sigma_A^2, \sigma_A^2)^\top$ gives

$$\theta_{AJ}^J = \frac{\sigma_A^2 \kappa_E^J \sigma_{AE}^2}{D_J}, \tag{23}$$

$$\theta_{AE}^J = \frac{\sigma_A^2 \kappa_J^J \sigma_{AJ}^2}{D_J}. \tag{24}$$

Symmetrically, for English investors,

$$\theta^E_{AJ} = \frac{\sigma^2_A \kappa^E_E \sigma^2_{AE}}{D_E}, \qquad \theta^E_{AE} = \frac{\sigma^2_A \kappa^E_J \sigma^2_{AJ}}{D_E}, \tag{25}$$

where

$$D_E \equiv \sigma^2_A(\kappa^E_J \sigma^2_{AJ} + \kappa^E_E \sigma^2_{AE}) + \kappa^E_E \kappa^E_J \sigma^2_{AJ} \sigma^2_{AE}. \tag{26}$$

The unaligned weights $\theta^i_{UJ}, \theta^i_{UE}$ follow directly from the diagonal $\Sigma^{(i)}_U$.

B.4 Equilibrium Prices and Return Slopes

Each investor type has mean–variance preferences over end–of–period wealth and can trade the risky asset and the risk–free asset at time 0. Let $a_i > 0$ denote the coefficient of absolute risk aversion for type i and P the time–0 price of the risky asset (per share).

Demands. Given s, investor i believes the conditional mean and variance of x are

$$\mathrm{E}_i[x \mid s] = \theta_i^\top s, \tag{27}$$

$$\mathrm{Var}_i(x \mid s) = \sigma^2_{x|s,i} = \mathrm{Var}(x) - \lambda^\top \Sigma_i^{-1} \lambda, \tag{28}$$

where $\sigma^2_{x|s,i} > 0$ is constant in s in the Gaussian–linear framework. Standard mean–variance calculations imply that investor i's optimal demand h_i (number of shares per unit wealth) is

$$h_i(s) = \frac{\mathrm{E}_i[x \mid s] - P}{a_i \mathrm{Var}_i(x \mid s)} = \lambda_i\big(\theta_i^\top s - P\big), \qquad \lambda_i \equiv \frac{1}{a_i \sigma^2_{x|s,i}} > 0. \tag{29}$$

Equilibrium price and price–implied weights. Let the aggregate mass of investors be normalized to one, so q_i is also the wealth share of type i. With net supply of the risky asset equal to zero (the supply level only affects the intercept, not the slope coefficients), market clearing requires $\sum_i q_i h_i(s) = 0$ for all realizations of s, i.e.,

$$\sum_{i \in \{J,E\}} q_i \lambda_i (\theta_i^\top s - P) = 0. \tag{30}$$

Solving for P yields

$$P = \beta^\top s, \qquad \beta_k \equiv \frac{\sum_i q_i \lambda_i \theta_{i,k}}{\sum_i q_i \lambda_i},\ k \in \{AE, AJ, UE, UJ\}. \tag{31}$$

Thus the equilibrium price is linear in the signal vector s, and the *price–implied weight* vector $\beta = (\beta_{AJ}, \beta_{AE}, \beta_{UJ}, \beta_{UE})^\top$ is a risk–tolerance–weighted average of the type–specific posterior weights. Investors with larger wealth shares q_i and greater risk tolerance λ_i have more influence on prices.

In the omnilingual benchmark without segmentation, there is a single investor type with $\Sigma_i = \Sigma$, so $\theta_i = \omega$ and $\beta = \omega$. In that case prices fully reflect available information.

Return regression and inefficiency. Define the realized excess return as $R \equiv x - P$. Combining $\widehat{x}^{\text{full}}(s) = \omega^\top s$ from (8) with $P = \beta^\top s$ from (31) and the payoff decomposition $x = \omega^\top s + \eta$ (where η is the projection error with $\mathrm{E}[\eta \mid s] = 0$) gives

$$R = x - P = (\omega - \beta)^\top s + \eta. \tag{32}$$

Consider the population regression of R on s with date fixed effects,

$$R = \alpha_t + \sum_{k \in \{AJ,AE,UJ,UE\}} \gamma_k s_k + \varepsilon, \tag{33}$$

as in equation (1) of the main text. In vector form this is $R = \alpha_t + \gamma^\top s + \varepsilon$, where $\gamma = (\gamma_{AJ}, \gamma_{AE}, \gamma_{UJ}, \gamma_{UE})^\top$.

Because s is exogenous and the model is linear–Gaussian, the population OLS coefficient vector is

$$\gamma = \Sigma^{-1}\mathrm{Cov}(s, R).$$

We have

$$\mathrm{Cov}(s, R) = \mathrm{Cov}\big(s, (\omega - \beta)^\top s + \eta\big) = \Sigma(\omega - \beta),$$

since $\mathrm{Cov}(s, \eta) = 0$. Hence

$$\gamma = \Sigma^{-1}\Sigma(\omega - \beta) = \omega - \beta. \tag{34}$$

Componentwise,

$$\gamma_k = \omega_k - \beta_k,$$

which is equation (2) in the main text. We interpret γ_k as a measure of inefficiency in how score k is reflected in prices. A value $\gamma_k \approx 0$ corresponds to efficient pricing of score k, while $\gamma_k > 0$ (< 0) corresponds to under–weighting (over–weighting) that score relative to the full–information benchmark.

B.5 Proofs

B.5.1 Proof of Prediction 1

We assume the baseline case, $\kappa_E^J, \kappa_J^E > 1$ and $\kappa_J^J, \kappa_E^E = 1$, that both types have market weight $q_i > 0$.

Proof. Consider first $k = UJ$. For an unaligned score, the omnilingual (full-information) weight is

$$\omega_{UJ} = \frac{\mathrm{Cov}(s_{UJ}, x)}{\mathrm{Var}(s_{UJ})}.$$

Under segmented processing, type i uses an inflated perceived variance

$$\mathrm{Var}_i(s_{UJ}) = \mathrm{Var}(s_{UJ}) + \delta_{i,UJ}, \qquad \delta_{i,UJ} \geq 0,$$

with $\delta_{i,UJ} = (\kappa_J^i - 1)\text{Var}(\varepsilon_{UJ})$. The type-$i$ posterior weight on s_{UJ} is

$$\theta_{i,UJ} = \frac{\text{Cov}(s_{UJ}, x)}{\text{Var}_i(s_{UJ})} = \frac{\text{Var}(s_{UJ})}{\text{Var}(s_{UJ}) + \delta_{i,UJ}}\, \omega_{UJ}.$$

The assumptions $\kappa_J^J = 1$ and $\kappa_J^E > 1$ imply $\delta_{J,UJ} = 0$ and $\delta_{E,UJ} > 0$, hence

$$\theta_{J,UJ} = \omega_{UJ}, \qquad 0 < \theta_{E,UJ} < \omega_{UJ},$$

where the strict inequalities use $\omega_{UJ} > 0$. Define

$$\phi_J = \frac{q_J \lambda_J}{q_J \lambda_J + q_E \lambda_E}, \qquad \phi_E = \frac{q_E \lambda_E}{q_J \lambda_J + q_E \lambda_E},$$

so $\phi_J, \phi_E > 0$ and $\phi_J + \phi_E = 1$. Then

$$\beta_{UJ} = \phi_J \theta_{J,UJ} + \phi_E \theta_{E,UJ} = \phi_J \omega_{UJ} + \phi_E \theta_{E,UJ} < \phi_J \omega_{UJ} + \phi_E \omega_{UJ} = \omega_{UJ}.$$

Therefore $\gamma_{UJ} = \omega_{UJ} - \beta_{UJ} > 0$.

The case $k = UE$ is symmetric. Here $\delta_{E,UE} = (\kappa_E^E - 1)\text{Var}(\varepsilon_{UE}) = 0$ and $\delta_{J,UE} = (\kappa_E^J - 1)\text{Var}(\varepsilon_{UE}) > 0$, implying $\theta_{E,UE} = \omega_{UE}$ and $0 < \theta_{J,UE} < \omega_{UE}$. The same convex-combination argument yields $\beta_{UE} < \omega_{UE}$ and hence $\gamma_{UE} > 0$. ■

B.5.2 Proof of Prediction 2

Assume again that $\kappa_E^J, \kappa_J^E > 1$ and $\kappa_J^J, \kappa_E^E = 1$ and that both types have market weight $q_i > 0$.

Proof. Let $a \equiv \sigma_A^2$, $j \equiv \sigma_{AJ}^2$, and $e \equiv \sigma_{AE}^2$. In the omnilingual benchmark, the aligned covariance block is

$$\Sigma_A = \begin{pmatrix} a + j & a \\ a & a + e \end{pmatrix}, \qquad \lambda_A = \begin{pmatrix} a \\ a \end{pmatrix}.$$

Hence the full-information (Bayesian) weight vector on (s_{AJ}, s_{AE}) is $\omega_A = \Sigma_A^{-1} \lambda_A$, and the total aligned weight is

$$\omega_{AJ} + \omega_{AE} = \frac{a(j + e)}{D_A}, \qquad D_A \equiv a(j + e) + je.$$

Under language-segmented processing with $\kappa_J^J = \kappa_E^E = 1$, Japanese investors inflate only the English aligned noise variance by κ_E^J:

$$\Sigma_A^{(J)} = \begin{pmatrix} a + j & a \\ a & a + \kappa_E^J e \end{pmatrix}, \qquad D_J \equiv a(j + \kappa_E^J e) + \kappa_E^J je,$$

so their total aligned posterior weight equals

$$\theta_{AJ}^J + \theta_{AE}^J = \frac{a(j + \kappa_E^J e)}{D_J}.$$

A direct comparison gives

$$(\omega_{AJ}+\omega_{AE})-(\theta^J_{AJ}+\theta^J_{AE})=a\left[\frac{j+e}{D_A}-\frac{j+\kappa^J_E e}{D_J}\right]=a\cdot\frac{j^2e(\kappa^J_E-1)}{D_AD_J}>0,$$

because $a,j,e>0$ and $\kappa^J_E>1$. Hence $\theta^J_{AJ}+\theta^J_{AE}<\omega_{AJ}+\omega_{AE}$.

Similarly, English investors inflate only the Japanese aligned noise variance by κ^E_J:

$$\Sigma^{(E)}_A=\begin{pmatrix}a+\kappa^E_J j & a\\ a & a+e\end{pmatrix},\qquad D_E\equiv a(\kappa^E_J j+e)+\kappa^E_J je,$$

so

$$\theta^E_{AJ}+\theta^E_{AE}=\frac{a(\kappa^E_J j+e)}{D_E},$$

and

$$(\omega_{AJ}+\omega_{AE})-(\theta^E_{AJ}+\theta^E_{AE})=a\cdot\frac{je^2(\kappa^E_J-1)}{D_AD_E}>0,$$

since $a,j,e>0$ and $\kappa^E_J>1$. Thus $\theta^E_{AJ}+\theta^E_{AE}<\omega_{AJ}+\omega_{AE}$.

In equilibrium, the price weight vector is a convex combination of type-specific weights: $\beta=\phi_J\theta^J+\phi_E\theta^E$ with $\phi_J,\phi_E>0$ and $\phi_J+\phi_E=1$. Therefore,

$$\beta_{AJ}+\beta_{AE}=\phi_J(\theta^J_{AJ}+\theta^J_{AE})+\phi_E(\theta^E_{AJ}+\theta^E_{AE})<\omega_{AJ}+\omega_{AE}.$$

Finally, by definition $\gamma_k=\omega_k-\beta_k$, so

$$\gamma_{AJ}+\gamma_{AE}=(\omega_{AJ}+\omega_{AE})-(\beta_{AJ}+\beta_{AE})>0.$$

■

Remark about $\gamma_{UJ}<0$. If $\kappa^J_J<1$ (Japanese investors are overconfident in native Japanese news) and $\kappa^E_J\geq 1$ (English investors do not overweight Japanese news) and Japanese investors have sufficient equilibrium price weight, the model can generate $\gamma_{UJ}<0$.

Specifically, if $\kappa^J_J<1$ and $\kappa^E_J\geq 1$, then $\theta_{J,UJ}>\omega_{UJ}\geq\theta_{E,UJ}$, so $\theta_{J,UJ}>\theta_{E,UJ}$ and

$$\beta_{UJ}>\omega_{UJ}\iff\phi(\theta_{J,UJ}-\theta_{E,UJ})>\omega_{UJ}-\theta_{E,UJ}\iff\phi>\frac{\omega_{UJ}-\theta_{E,UJ}}{\theta_{J,UJ}-\theta_{E,UJ}}\in(0,1).$$

Notably, $\gamma_{UJ}<0$ does not in general imply $\gamma_{AJ}<0$ because s_{AJ} is not a stand-alone signal: it is one of two noisy measurements of the same aligned payoff component A, with the other being the aligned English score s_{AE}. By contrast, $\gamma_{AJ}<0$ requires $\beta_{AJ}>\omega_{AJ}$ with $\beta_{AJ}=\phi\theta_{J,AJ}+(1-\phi)\theta_{E,AJ}$, and the aligned posterior weight $\theta_{i,AJ}$ depends on how type i also perceives the aligned English signal. In the two-signal aligned block,

$$\theta_{i,AJ}=\frac{\sigma^2_A\,\kappa^i_E\sigma^2_{AE}}{\sigma^2_A(\kappa^i_J\sigma^2_{AJ}+\kappa^i_E\sigma^2_{AE})+\kappa^i_J\kappa^i_E\sigma^2_{AJ}\sigma^2_{AE}},$$

so the threshold for $\gamma_{AJ} < 0$,

$$\phi^{\star}_{AJ} \equiv \frac{\omega_{AJ} - \theta_{E,AJ}}{\theta_{J,AJ} - \theta_{E,AJ}},$$

depends on (κ^J_E, κ^E_E) and the aligned covariance structure, not just on the unaligned Japanese multipliers (κ^J_J, κ^E_J). Hence one can have $\phi^{\star}_{UJ} < \phi < \phi^{\star}_{AJ}$, implying $\gamma_{UJ} < 0$ but $\gamma_{AJ} \geq 0$: overreaction to Japanese-only details need not imply overreaction to aligned Japanese news because aligned English coverage provides corroborating information about A that can discipline the pricing of s_{AJ}. Extending the model further, e.g., by allowing signal-specific native multiplies (e.g., κ^J_{UJ}), can generate even richer behavior.