CrossAlpha: An Annual-Report Benchmark for Cross-Market Factor Research with LLM Agents

Qian Wang
National University of Singapore
qiansoc@nus.edu.sg

Zhongyi Tong
MS Capital
zhongyi.tong@mscapital.sg

Nuo Chen
National University of Singapore
nuochen@u.nus.edu

Zhaomin Wu
National University of Singapore
zhaomin@nus.edu.sg

Bingsheng He
National University of Singapore
dcsheb@nus.edu.sg

Abstract

Cross-market factor research studies whether firm-level signals from one or more markets can predict returns in a target market, but existing public benchmarks do not support cross-market disclosure-to-return evaluation. Building such a benchmark is challenging because filings differ across languages and regulatory systems, disclosure-derived similarity can be biased by common reporting components, and cross-market signals must be evaluated under feasible trading-time alignment. We introduce **CrossAlpha**, a public annual-report benchmark for cross-market factor research. CrossAlpha addresses these challenges through three corresponding components: *Disclosure Distillation*, which standardises heterogeneous filings into ten-category English business descriptions; *Residual Schema Graph Construction*, which builds PCA-whitened cross-market firm-pair scores from schema-level disclosures; and *Timing-Aligned Evaluation*, which pairs the graph with 11 years of daily OHLCV data to construct forward-return labels under feasible cross-market execution protocols. CrossAlpha covers 3,587 firms and about 10,700 firm-year reports from the United States, Japan, Taiwan, South Korea, and Hong Kong, and releases about 19M directed firm-pair scores. In experiments, disclosure-derived cross-market peers outperform domestic text, industry-code, and return-correlation peers in the US-to-Japan setting (ICIR 0.39 versus 0.07–0.18), and cross-market sources beat the domestic text baseline in most target markets. CrossAlpha offers a publicly released, reusable, return-grounded benchmark for cross-market financial NLP.

1 Introduction

In trading, a factor is a firm-level score used to rank stocks: a useful factor assigns higher scores to firms that subsequently earn higher market-relative returns (Vuolteenaho, 2002). Cross-market factor research asks whether signals from a *source* market can rank stocks in a different *target* market (Menzly and Ozbas, 2010; Ma, 2019). A target firm receives a high factor value when economically linked source-market firms have recently earned strong market-relative returns. Annual reports are ideal for this construction because they describe counterparties, technologies, and end-market exposures *before* those relationships are fully priced into cross-market stock co-movement. As shown in Figure 1, an AI demand shock at NVIDIA can be translated into high factor scores for Asian semiconductor firms whose filings reveal related business exposures. Extracting cross-market alpha therefore presents a vital challenge for financial natural language processing (NLP): it requires mapping heterogeneous disclosures into a structural firm graph to capture these global return spillovers.

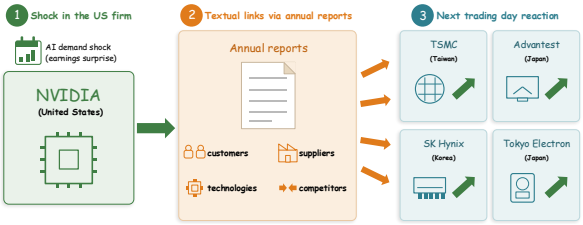


Figure 1: Cross-market information flow. A return shock to a source firm transmits to foreign targets via relationships embedded in annual reports.

However, existing resources address only isolated parts of this problem. Financial NLP benchmarks provide standardized evaluation for language understanding and retrieval, but they rarely test whether text-derived firm links generate economically meaningful, forward-looking return predictions (Shah et al., 2022; Jørgensen et al., 2023; Liu et al., 2024; Tang and Yang, 2025). Text-based

github.com/Persdre/crossalpha

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 19312–19341
October 24-29, 2026 ©2026 Association for Computational Linguistics

peer methods derive useful firm similarity from filings, but they are mostly developed within a single market, ignoring global supply chains (Hoberg and Phillips, 2016; Chung and Tanaka-Ishii, 2023; Takayanagi et al., 2024; Jagrič and Herman, 2026). Meanwhile, cross-market asset-pricing studies do evaluate return spillovers, but they typically rely on backward-looking price co-movement or statistically selected links rather than reusable, firm-level text (Lin et al., 2025; Bukhari et al., 2025; Liu et al., 2026). Thus, the field lacks a public benchmark that jointly connects multi-market disclosure text, dense economic links, aligned prices, and return-grounded factor evaluation.

To address these gaps, we introduce **CrossAlpha**, a public cross-market annual-report benchmark for cross-market factor research. CrossAlpha standardises the problem of turning long, multilingual filings into comparable firm representations, and evaluates the resulting cross-market links by the returns they predict. Because capturing meaningful cross-border information flow requires economies with large, liquid traded universes and deep integration through global tech and manufacturing supply chains, CrossAlpha focuses on five heavily linked markets that meet these criteria: the United States, Japan, Taiwan, South Korea, and Hong Kong.

Building a useful benchmark across these markets requires overcoming challenges at three distinct layers. First, at the *disclosure layer*, raw annual reports suffer from severe filing heterogeneity across different languages and regulatory systems. Second, at the *firm-link layer*, extracting genuine economic relationships is difficult because raw textual similarity is often biased by shared reporting-style components and broad industry jargon. Third, at the *evaluation layer*, cross-market return prediction must respect strict temporal boundaries; mismatched time zones and trading calendars can easily introduce look-ahead bias if execution lags are not rigorously handled.

To address these challenges, we introduce a three-stage construction and evaluation pipeline. To resolve filing heterogeneity, an LLM-based *Disclosure Distillation* stage converts the diverse annual reports into a shared ten-category English business schema. To mitigate similarity bias, a *Residual Schema Graph Construction* stage applies PCA whitening to the schema embeddings, removing common reporting components to yield a dense, directed cross-market graph. Finally, to ensure rigorous backtesting, a *Timing-Aligned Evaluation* protocol aligns the resulting semantic graph with 11 years of daily price data, supporting both monthly peer-momentum factors and event-conditioned evaluation with an LLM filter under feasible execution lags.

Our contributions are summarized as follows:

- We introduce CrossAlpha, an extensive public benchmark connecting standardized annual-report text, dense semantic firm-to-firm graphs, and 11 years of aligned daily prices across five major equity markets: the US, Japan, Taiwan, South Korea, and Hong Kong. It lays the groundwork for cross-market factor research by utilizing authentic economic relationships extracted from public corporate disclosures.

- CrossAlpha consists of a reproducible Distiller–Graph–Linker pipeline and a market-aware evaluation protocol. This inclusive benchmark captures not only static structural firm links through monthly peer-momentum factors but also dynamic market reactions through high-frequency event-conditioned tasks, providing a more holistic view of global return spillovers.

- Through detailed experiments, we demonstrate that disclosure-derived links markedly enhance cross-market return prediction. First, cross-market text peers outperform domestic and non-text baselines, achieving up to a five-fold improvement in predictive skill (e.g., an ICIR of 0.39 in the US-to-Japan setting). Second, source selection reveals a directed *Information Geography* that consistently yields positive long–short factors. Finally, using an LLM filter on graph-retrieved neighbors improves event-driven spillover trading, nearly doubling the portfolio Sharpe ratio.

2 Related Work

We discuss the most related work here and leave more details in Appendix D.

Comparison with Existing Resources. Table 1 separates the related work by the capabilities needed for a reusable cross-market factor benchmark. TNIC (Hoberg and Phillips, 2016) provides public 10-K-based firm links, but it is a single-market peer network rather than a return-prediction study. S-Peer (Chung and Tanaka-Ishii, 2023), MD&A Topic (Zhang et al., 2026), ECC

Table 1: Capability comparison for cross-market factor research. CrossAlpha is the only public resource that combines multi-market annual-report text, firm links, aligned prices, and a reusable return-prediction benchmark.

Resource / Method	Data					Evaluation	
	# Markets	Firm text	Firm links	Prices	Public	Cross-market	Return evaluation
TNIC (Hoberg and Phillips, 2016)	1	✓	✓	✗	✓	✗	✗
S-Peer (Chung and Tanaka-Ishii, 2023)	1	✓	✓	✓	✗	✗	✓
MD&A Topic (Zhang et al., 2026)	1	✓	✓	✓	✗	✗	✓
ECC Topic (Jin, 2024)	1	✓	✓	✓	✗	✗	✓
LLM-Augmented Semantic Network (Huang et al., 2026)	1	✓	✓	✓	✗	✗	✓
HybridFG (Du et al., 2025)	1	✗	✓	✓	✗	✗	✓
US–CN BG (Liu et al., 2026)	2	✗	✓	✓	✗	✓	✓
CSPO (Lin et al., 2025)	2	✗	✗	✓	✗	✓	✓
II-GNN (Bukhari et al., 2025)	5	✗	✗	✓	✗	✓	✓
CrossAlpha (ours)	**5**	✓	✓	✓	✓	✓	✓

Topic (Jin, 2024), and LLM-Augmented Semantic Network (Huang et al., 2026) evaluate text-derived firm links through peer-return or cross-stock predictability, but they remain single-market studies and do not release a reusable cross-market benchmark. HybridFG (Du et al., 2025), US–CN BG (Liu et al., 2026), CSPO (Lin et al., 2025), and II-GNN (Bukhari et al., 2025) provide return-prediction evaluations using supply-chain, price, or learned trading-signal graphs, but not a public annual-report semantic graph with reusable disclosure text, aligned prices, and factor protocols. CrossAlpha is designed to fill exactly this missing intersection: public firm text, public firm links, aligned prices, explicit cross-market evaluation, and a reusable return-prediction protocol.

3 CrossAlpha Construction

We build CrossAlpha with two construction techniques, followed by timing-aligned return evaluation, as shown in Figure 2. *Disclosure Distillation* addresses filing heterogeneity by converting annual reports from five disclosure systems into a shared ten-category English business schema (Sections 3.1 and 3.2; Appendix G). *Residual Schema Graph Construction* addresses common-component-biased similarity by embedding schema fields, applying PCA whitening to remove shared reporting-style components, and aggregating category-level residual similarities into a directed cross-market graph (Section 3.3; Appendix C.11). The remaining timing challenge is handled by *Timing-Aligned Evaluation* in the return-prediction protocols, where the graph is aligned with daily prices and tested under feasible execution lags.

3.1 Data Collection

Textual Filings. CrossAlpha covers five economically significant and tightly connected equity markets: the United States (Russell 1000, SEC EDGAR), Japan (TOPIX 500, EDINET), Taiwan (TWSE, MOPS), South Korea (KOSPI, DART), and Hong Kong (HK Main Board). These markets provide large listed-firm universes and dense US–Asia and intra-Asia trade, technology, and supply-chain linkages, making them a useful testbed for cross-market factor research. The firm universe consists of recent constituents of these indices or boards that have a 10-K-equivalent annual filing, so the benchmark describes a large, liquid listed universe. The official disclosure portals and their usage-policy links are listed in Appendix A.

Market Data. To backtest the factors we extract from the cross-market dataset, we collect daily OHLCV (open, high, low, adjusted close, and volume) from Yahoo Finance[1] for January 2015–May 2026 across all five markets, together with historical shares outstanding for market-capitalisation neutralization.

3.2 Disclosure Distillation

We design the Distiller schema to convert heterogeneous annual reports into comparable business-exposure representations for cross-market firm-link construction. The schema satisfies two requirements: each category should capture a cross-firm channel that prior return-predictability work finds informative, and each category should be recoverable from mandatory annual filings across all five regulatory systems. The resulting ten-category English schema spans products and technologies, customers and supply chains, geographic and finan-

[1] https://finance.yahoo.com

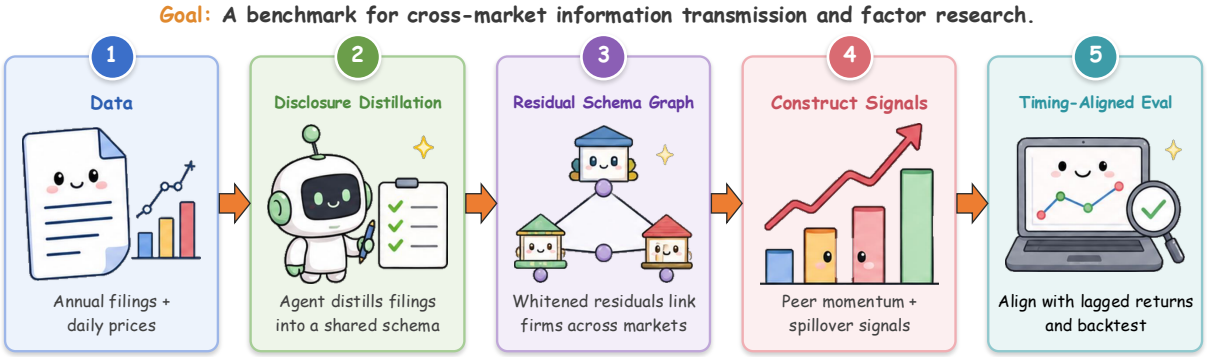


Figure 2: CrossAlpha Overview. The framework standardises annual reports from multiple equity markets through Disclosure Distillation, constructs a Residual Schema Graph from the resulting schema-field representations, and aligns the graph with price data for timing-aware return-prediction benchmarks.

cial exposure, and business model, strategy, and competition. Filings stay in their original language; the Distiller summarises their firm-level content into English, which serves as a shared pivot so that cross-market similarity is not driven by language differences. This design makes downstream similarity depend on comparable firm-level business content rather than market-specific filing formats. The category rationale and cross-regulator grounding are reported in Table 31 and Appendix G.

Given LLMs' capability to read financial text and extract structured information reliably (Lopez-Lira and Tang, 2023; Kim et al., 2024), we instantiate the schema with an LLM-based Distiller: a single GPT-4.1 (OpenAI, 2025a) pass per filing reads the document end-to-end and emits the standardised fields; the full prompt is given in Appendix F.1.

3.3 Residual Schema Graph Construction

Residual Schema Encoding. We embed each standardised category text with OpenAI's text-embedding-3-large (OpenAI, 2024) to obtain category-level firm representations. However, even after standardisation, these embeddings retain strong common components: all summaries are written in a similar business-description style, which can pull unrelated firms closer in vector space and inflate raw cosine similarities (Su et al., 2021). Since cross-market peer discovery should depend on firm-specific business exposure rather than shared reporting style, we apply PCA whitening to the category vectors before computing similarities (Kessy et al., 2018; Su et al., 2021). This transformation suppresses dominant common components and balances variance across embedding directions, making the remaining similarities more reflective of firm-specific schema content. We refer to the resulting PCA-whitened category vectors as residual schema encodings.

Directed Graph Construction. Using the residual schema encodings, CrossAlpha connects economically similar firms across markets into a dense directed graph, which we call the Residual Schema Graph:

$$\mathcal{G}_{\text{text}} = (V, E_{\text{text}}), \qquad V = \bigcup_{m=1}^{K} V_{\mathcal{M}_m}, \qquad \mathcal{M} \in \{\text{US}, \text{JP}, \text{TW}, \text{KR}, \text{HK}\}, \tag{1}$$

where nodes are listed firms partitioned by home market and edges represent firm-to-firm residual schema similarities. For an ordered source market $\mathcal{M}_b$ and target market $\mathcal{M}_a$, the similarity between a target firm $i \in \mathcal{M}_a$ and a source firm $j \in \mathcal{M}_b$ is the equal-weighted average of their ten category-level cosine similarities:

$$s_{\text{text}}(i, j) = \frac{1}{10} \sum_{c=1}^{10} \cos\left(\mathbf{z}_i^{(c)}, \mathbf{z}_j^{(c)}\right), \tag{2}$$

where $\mathbf{z}_i^{(c)} \in \mathbb{R}^{128}$ is the residual schema encoding of firm i's category-c text. The graph is dense and weighted: every firm pair receives a score, including same-market pairs, which form the diagonal of the market-pair matrix and the domestic-text baselines. Downstream tasks then concentrate weight on the top-ranked peers through the percentile-rank sigmoid of Equation (10). This construction yields a dense firm-to-firm peer layer across all 25 directed market pairs.

3.4 Dataset Statistics

Table 2 highlights three properties of CrossAlpha. First, the benchmark provides broad market coverage: it spans 3,587 listed firms across the United

States, Japan, Taiwan, South Korea, and Hong Kong, giving each target firm a large cross-market peer universe. Second, the released artifacts are aligned across modalities: the 10,700 firm-year schema records are paired with category-level representations, a dense firm-pair graph, and daily OHLCV data. Third, the graph layer expands the filing layer into about 19M directed firm-pair scores, providing the evaluation surface for cross-market peer selection and return prediction. Construction costs are in Appendix E.2.

Table 2: CrossAlpha dataset summary.

Filing universe		
Market	**Universe**	**# of Firms**
United States	Russell 1000	880
Japan	TOPIX 500	453
Taiwan	TWSE	1,040
South Korea	KOSPI	612
Hong Kong	Main board	602
Total	**5 equity markets**	**3,587**
Benchmark components		
Layer	**Unit**	**Scale**
Schema text	ten-category firm-year records	10,700
Residual schema encodings	firm-category vectors	≈107K
Residual Schema Graph	directed firm-pair scores	≈19M
Daily prices	firm-day OHLCV records	≈10M
Total	**schema, encoding, graph, and price records**	**≈29.1M**

4 Timing-aligned Evaluation Protocol

CrossAlpha serves not only as a textual and structural resource but also provides a standardized, return-grounded evaluation protocol. To isolate the predictive value of disclosure links, this protocol fixes the portfolio construction, varying only the source of firm links or the evaluation horizon.

Universe, portfolio, and timing. The released OHLCV panel spans January 2015–May 2026. The monthly factor tasks use January 2015–December 2025 with a strict one-month return lag: peer returns through month $M-1$ are used to form the month-M target-stock portfolio. Crucially, the backtests use the released FY2024 text-similarity graph as a fixed benchmark input; return information enters solely through the lagged peer returns. For each source–target market pair, the traded universe is strictly the target market. Source-market firms only define peer sets and supply lagged returns, while target-market stocks are ranked into quintiles and held as an equal-weighted Q5–Q1 long–short book, as shown in Figure 3. While returns are strictly lagged to prevent price leakage, applying the static FY2024 graph to the historical panel evaluates the structural return-relevance of these textual links, rather than simulating an out-of-sample trading strategy (which would require point-in-time dynamic graphs).

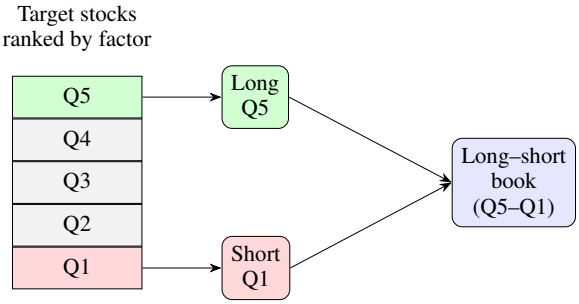


Figure 3: Portfolio sort. Each month, target-market stocks are ranked by factor value and split into five equal-count quintiles. The long–short (LS) book buys the top quintile (Q5) and sells the bottom (Q1).

Factor Construction. For a target stock $i \in \mathcal{M}_a$ and source market $\mathcal{M}_b$, the factor is the similarity-weighted average of source peers' sector-relative returns,

$$f_i^{\text{text}}(t) = \frac{\sum_{j \in \mathcal{M}_b} \alpha_{ij}^{\text{text}} \cdot r_j^{(L)}(t)}{\sum_{j \in \mathcal{M}_b} \alpha_{ij}^{\text{text}}}, \qquad (3)$$

where $\alpha_{ij}^{\text{text}}$ is the sigmoid top-1% similarity weight (Equation (10)) and $r_j^{(L)}(t)$ is peer j's L-month sector-relative return. Baselines reuse this identical pipeline, altering only the peer definition; source-selection variants change only which source market supplies the peers.

Metrics. We report six equity-factor metrics: (1) IC and (2) ICIR measure rank-predictive skill (Grinold and Kahn, 2000), while the long–short (3) Sharpe (Sharpe, 1994), (4) MaxDD, (5) Ret, and (6) CumRet summarize the portfolio after a 2 bp one-way transaction cost. IC is the monthly Spearman rank correlation between factor scores and next-month returns. ICIR is the annualized mean-to-volatility ratio of that monthly IC series, following standard firm-link return-predictability literature (Cohen and Frazzini, 2008; Menzly and Ozbas, 2010). Full metric definitions and equations are provided in Appendix C.1 and Equations (4) and (5).

Inference. Because monthly IC series are autocorrelated and the market-pair analysis tests many cells at once, we compute Newey–West standard errors (Newey and West, 1987) and correct for multiple comparisons (Benjamini and Hochberg, 1995; Holm, 1979); the resulting statistics are reported in Appendix C.5.

Neutralization. For cross-market source comparisons—including source selection, Cross-Market Gain, and Information-Geography—we report signed ICIR after GICS-sector and size neutralization. We retain the sign to explicitly capture reverse predictability (negative values). This source-comparison metric aligns with recent relational benchmarks (Bukhari et al., 2025; Du et al., 2025), ensuring that we measure purely relational, within-sector and within-size skill rather than a latent sector or size tilt, with details in Appendix C.2.

5 Experiments

Using the timing-aligned protocol in Section 4, we evaluate CrossAlpha through three benchmark questions: **RQ1:** Do peers defined by cross-market filing text predict returns better than peers from industry codes, price correlation, or a single market? **RQ2:** Where should those peers come from, and is one source market best for every target or does the best source differ by target? **RQ3:** Beyond the monthly factor, does the same peer graph support event-conditioned daily spillover prediction?

5.1 RQ1: Cross-Market Text vs. Single-Market and Non-Text Peers

Approach. We fix one source–target market pair: target Japan TOPIX 500, and source US Russell 1000. The US offers a broad, globally connected source universe, and Japan is a large, fully covered target with the lowest single-sector concentration among the Asian markets, so the comparison spans many sectors; Appendix C.4 gives the market statistics. All methods share the identical monthly peer-momentum pipeline and differ only in how peers are defined. We compare CrossAlpha's cross-market text embedding against three alternatives: (i) a single-market text baseline (JP-to-JP domestic peers, identical pipeline), which tests whether cross-market text is needed at all; (ii) GICS sector equal-weight peers; and (iii) 252-day rolling return correlation, the standard non-text peer definition. Full baseline settings are in Appendix C.3.

Table 3 reports the comparison and Figure 4 plots the cumulative long-short returns. We have two findings as follows:

Table 3: US-to-JP comparison. All rows share the same monthly pipeline and differ only in the peer definition.

Peer definition	IC	ICIR	Sharpe	MaxDD	Ret	CumRet
GICS sector	0.006	0.18	0.25	-14.3%	1.1%	11.6%
Return correlation	0.004	0.11	0.23	-8.4%	1.0%	10.7%
JP-to-JP text	0.003	0.07	0.05	-9.5%	0.3%	3.0%
US-to-JP text	0.009	0.39	0.39	-6.8%	1.2%	13.2%

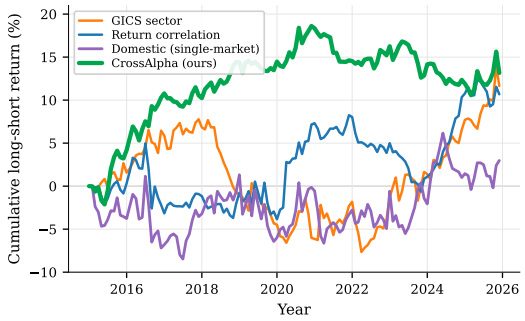


Figure 4: Cumulative long-short return on US-to-JP under the identical monthly pipeline, with methods differing only in the peer definition.

Cross-market text beats the single-market baseline. Building peers from domestic Japanese filings barely helps. JP-to-JP text peers reach an ICIR of only 0.07, so same-market text similarity carries almost no predictive signal. Drawing the same peers from the US instead lifts the ICIR to 0.39. We attribute this gap to where the information originates rather than to the text itself: investors price firms inside their own market almost immediately, leaving no time gap for a domestic peer to exploit, whereas a US disclosure is reflected in US prices first and reaches the linked Japanese firm only later, and the factor profits from that delay (Menzly and Ozbas, 2010). Figure 4 makes the contrast visible: the cross-market portfolio compounds steadily and stays above every baseline, while the domestic portfolio drifts around zero.

Text also beats the non-text peer definitions, which break cross-market. GICS sector matching reaches an ICIR of 0.18 and 252-day return correlation 0.11, against 0.39 for text, a two- to three-fold gap that also shows up in risk (Sharpe 0.39, max drawdown -6.8% versus GICS's -14.3% as shown in Table 3). We attribute this to what each definition measures: a GICS label is too coarse to say whether a US and a Japanese firm are actually each other's customer, supplier, or competitor, while re-

turn correlation is backward-looking and captures the result of information flow rather than its cause. Business-description text is more direct in this setting because it can identify firm-level economic links before the linked returns are observed.

The result holds up under stricter controls. Three concerns could explain the gap without any text-derived link, and we test each in the sector- and size-neutralized US-to-JP task of Appendix C.2, where the released graph reaches an ICIR of 0.526. The first is look-ahead: the released graph is built from FY2024 filings, so for earlier months it may encode relationships that did not yet exist. Rebuilding the US side from FY2023 filings and evaluating only after those filings were available still gives an ICIR of 0.37, compared with 0.57 for the released graph over the same window, and older FY2022 descriptions decay to 0.20; Appendix C.6 reports the full sensitivity. The second is that large US and Japanese firms simply co-move. Randomly reassigning US peers within the same sector, size, and liquidity group collapses the ICIR to -0.03, while dropping the most-liquid US decile leaves it at 0.51, so the identity of the linked firm is what matters; Appendix C.7 adds the size and liquidity subgroups. The third is that a price-based cross-market model would do as well. Controlled implementations of the relation mechanisms in II-GNN (Bukhari et al., 2025) and CSPO (Lin et al., 2025) on the same inputs reach ICIRs of 0.05 or below, against 0.55 for CrossAlpha over the same months, as detailed in Appendix C.8. Finally, injecting extraction noise at the worst-language audit rate of 18% retains 93% of the clean signal, as shown in Appendix C.10.

5.2 RQ2: Source Selection and Directed Information Geography

Because Section 5.1 shows in the US-to-JP setting that cross-market text beats single-market and non-text peer definitions, we next ask whether that result is general for other markets. We therefore scale the same monthly factor construction to all five markets and ask which source market gives the strongest factor for each target and how directional those cross-market links are.

Approach. We extend the single US-to-JP source–target pair to every ordered pair of markets. For each source s and target t in {US, JP, TW, KR, HK} we build the same peer-momentum factor as in Section 5.1 and record signed annualised ICIR as a compact predictive statistic for comparing source-market choices. These scores form a 5-by-5 matrix $\mathbf{G}$ that we call the *Cross-Market Information-Geography Matrix*, where $\mathbf{G}_{st}$ is how strongly source s predicts target t. Reading a column gives each target's best single cross-market source, and comparing $\mathbf{G}_{st}$ with $\mathbf{G}_{ts}$ gives the direction of information flow.

Table 4: Cross-market generalization: US-to-T peer-momentum factors across four Asian targets.

Market	LB	IC	ICIR	Sharpe	MaxDD	Ret	CumRet
JP	6mo	0.008	0.31	0.30	-8.2%	1.0%	10.8%
JP	12mo	0.009	0.39	0.39	-6.8%	1.2%	13.2%
TW	6mo	0.008	0.41	0.71	-5.9%	2.4%	26.6%
TW	12mo	0.009	0.48	0.91	-3.0%	2.9%	32.1%
KR	6mo	0.009	0.46	0.49	-8.3%	2.4%	25.6%
KR	12mo	0.012	0.61	0.64	-4.3%	5.6%	60.8%
HK	6mo	0.015	0.66	0.98	-7.1%	5.0%	54.1%
HK	12mo	0.013	0.52	0.88	-6.4%	4.7%	50.7%

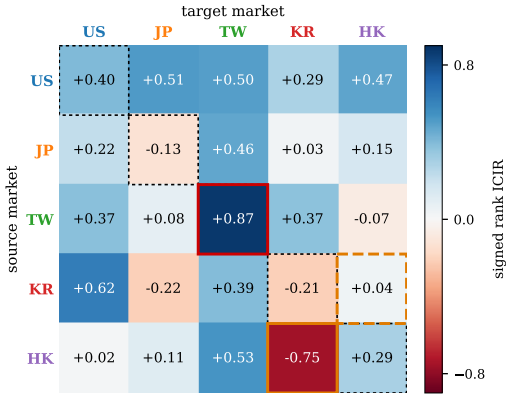


Figure 5: Cross-Market Information-Geography Matrix $\mathbf{G}$. (s, t) is the signed ICIR of the source-s-to-target-t factor; diagonal cells are same-market pairs scored by the same dense graph.

Analyzing this Information-Geography Matrix reveals a complex network of global predictability. Specifically, we draw four key insights regarding how and where alpha signals propagate:

CrossAlpha finds positive cross-market factors across all five targets. As shown in Table 4, the benchmark turns annual-report links into positive cross-market factors. After selecting the best non-domestic source for each target, the long–short factor is positive in every market in Table 5, with annualised returns from 0.3% to 4.0% and Sharpe ratios up to 0.82. The fixed US-source experiment tells the same story from a stricter angle: even when using only US peers as the source, the factor performs strongly across Asian target markets.

Source selection reveals target-specific information geography. The strongest non-domestic

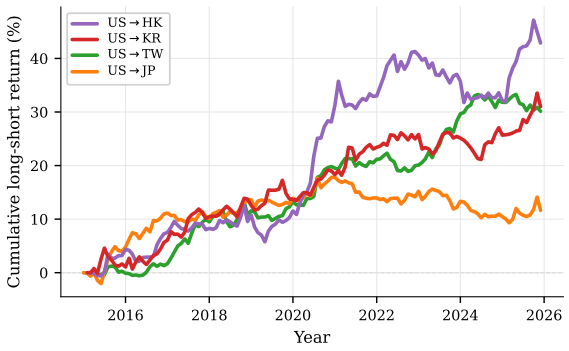


Figure 6: Cross-market generalization across four Asian targets.

Table 5: Cross-market performance for each target market. The ICIR block compares the cross-market signal from the source matrix with the same-market baseline, with Gain defined as Cross-market minus Domestic.

Target	ICIR			Portfolio		
	Cross-market	Domestic	Gain	Sharpe	MaxDD	Ret
US	0.62	0.40	0.22	0.19	-5.8%	0.4%
JP	0.51	-0.13	0.64	0.06	-11.0%	0.3%
TW	0.53	0.87	-0.34	0.82	-5.3%	2.5%
KR	0.37	-0.21	0.58	0.06	-26.2%	0.3%
HK	0.47	0.29	0.18	0.82	-5.4%	4.0%

source changes with the target market: Korea is best for the US, the US is best for Japan and Hong Kong, Hong Kong is best for Taiwan, and Taiwan is best for Korea, as shown in Figure 5. Four different source markets appear across five targets, so CrossAlpha exposes information geography that a fixed-source benchmark would miss.

Cross-market links provide signal beyond domestic peers. The Gain column in Table 5 compares the cross-market signal with the same-market text baseline. It is positive for four of five targets, showing that cross-border annual-report links add information beyond domestic peers. Taiwan is the exception: its domestic ICIR remains higher than any cross-market source, consistent with a tightly linked local semiconductor cluster whose most informative peers are still at home (Wang and Chiu, 2014).

Information flow is directional rather than symmetric. Reversing source and target often changes both the strength and the sign of predictability; for example, Hong Kong predicts Taiwan, while Taiwan-to-Hong Kong is negative. This asymmetry shows that CrossAlpha is not merely recovering undirected business similarity, but identifies where return-relevant information appears first.

Individual cells should be read with care. Because the matrix tests 25 market pairs at once, we correct for multiple comparisons. After false-discovery correction only the strongest cell, TW-to-TW, remains significant on its own, as Appendix C.5 reports, so we read the matrix as a map of where the useful source markets lie.

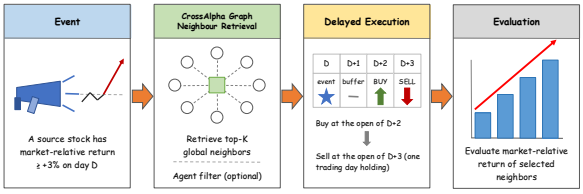


Figure 7: Event-driven evaluation pipeline.

Table 6: Event-conditioned LLM-filter evaluation setup.

Component	Definition
Events	Source-stock day D with market-relative close-to-close return above +3%; 25 source stocks fixed before the test window (top-5 by 2023-12-31 market cap in each market; Appendix C.9).
Graph baskets	For each event, rank the source stock's global embedding neighbours and evaluate equal-weight top-K baskets for $K \in \{10, 20, 30, 60\}$.
LLM filter	The LLM filter labels the top-60 graph neighbours as {co-mover, inverse, unrelated}; *Graph + LLM filter* keeps labelled co-movers and cannot add new firms, making the LLM an auditable filter rather than a generator.
Random controls	*Market random* matches the graph basket's market mix; *Market+sector random* matches both market and GICS sector. Each random bar averages 200 fixed-seed draws.
Timing	Strict $t+2$ open-to-open execution: buy at $D+2$ open and sell at $D+3$ open.
Bias control	All events are after GPT-5's 2024-09-30 knowledge cutoff[2]; full construction details are in Appendix C.9.

5.3 RQ3: Event-Conditioned Daily Spillover Evaluation with an LLM Filter

Sections 5.1 and 5.2 evaluate CrossAlpha as a monthly peer-momentum benchmark. We now ask whether the same graph supports a higher-frequency, event-conditioned trading task: after a large idiosyncratic move in a source firm, do its graph neighbours earn abnormal market-relative returns in subsequent daily trading?

Approach. As shown in Figure 7, we evaluate a post-shock neighbour-retrieval task: after a large source-stock move, the graph retrieves top-K global neighbours, and the GPT-5 LLM filter narrows that graph shortlist into event-specific co-movers. The LLM filter cannot propose new firms, so the test isolates whether LLM filtering improves graph retrieval. We use strict $t+2$ open-to-open execution so all source–target time-zone orderings are feasible. Detailed setup is in Table 6. From the results in Table 7, we have the following insights:

Table 7: Event-conditioned cross-market transmission across basket sizes K on 606 events. Graph is the CrossAlpha graph basket; Graph+LLM filter applies the GPT-5 co-mover filter; the random controls match market or market+sector composition. Per-event t-statistics are in Table 21.

K	Basket	Event Ret	Sharpe	CumRet
10	Graph	0.107%	1.32	76.5%
	Graph+LLM filter	0.157%	1.84	136.5%
	Market rand.	-0.000%	-0.02	-0.3%
	Market+sector rand.	-0.005%	-0.11	-3.3%
20	Graph	0.051%	0.81	30.0%
	Graph+LLM filter	0.108%	1.50	80.8%
	Market rand.	-0.000%	-0.02	-0.2%
	Market+sector rand.	0.008%	0.21	4.6%
30	Graph	0.073%	1.36	50.2%
	Graph+LLM filter	0.102%	1.45	74.8%
	Market rand.	-0.003%	-0.11	-1.6%
	Market+sector rand.	0.014%	0.44	8.7%
60	Graph	0.052%	1.41	35.1%
	Graph+LLM filter	0.104%	1.49	77.3%
	Market rand.	0.000%	0.01	0.1%
	Market+sector rand.	0.015%	0.64	9.3%

Graph signals extend to daily event spillovers, outperforming structural controls. The monthly graph signal carries over to the daily event task: as shown in Table 7, graph neighbours earn positive market-relative event returns at every basket size, with mean event returns of 0.05%–0.11%, annualised Sharpe ratios up to 1.41, and cumulative compound returns up to 76.5%. Both market-matched and market-and-sector-matched random baskets stay near zero or much weaker on the same metrics. This shows that the text-derived ranking adds signal beyond coarse market or sector allocation, rather than merely selecting firms from the right region or industry.

LLM filtering strengthens the fixed graph shortlist. The GPT-5 LLM filter acts strictly as a filter, classifying the retrieved top-K neighbours into co-movers, inverse exposures, or unrelated firms without adding new names. Despite this restricted role, the LLM filter improves all return and risk-adjusted metrics across basket sizes in Table 7. At K=20, the annualized Sharpe ratio rises from 0.81 to 1.50, and the cumulative compound return jumps from 30.0% to 80.8%. The improvement persists when the filter is swapped, as shown in Table 23: GPT-5.5 reaches a Sharpe ratio of 1.57 and the open-weight Qwen3-235B reaches 1.40, whereas a same-sector rule without an LLM stays at 0.38. The NVIDIA case illustrates this filtering mechanism in practice. While Nintendo is structurally retrieved due to shared GPU and gaming vocabulary, the LLM filter recognizes that the specific event concerns data-center AI rather than console demand and removes it. This reasoning proves accurate, as Nintendo subsequently realizes a -2.83% market-relative return under the t+2 execution rule (see Appendix C.9 for details).

[2]Knowledge cutoff of OpenAI's GPT-5 https://platform.openai.com/docs/models/gpt-5.

6 Conclusion

CrossAlpha is a public annual-report dataset for studying cross-market firm links through downstream return prediction. By aligning standardized disclosures, dense semantic firm graphs, and 11 years of daily prices across five equity markets, CrossAlpha moves financial-NLP evaluation beyond document-level text quality toward market-grounded firm-link reasoning. Our experiments show that disclosure-derived cross-market links capture predictive relationships missed by domestic text, industry labels, and price-only peer definitions, and that the useful source market is directed and target-specific rather than uniformly US-led. The same graph also supports event-conditioned spillover evaluation, where LLM-filtered graph neighbours improve daily post-event baskets. We release CrossAlpha as a shared benchmark bridging machine learning and quantitative finance.

Limitations

CrossAlpha is a benchmark for disclosure-grounded cross-market return prediction on large, liquid listed firms, and its annual-report graph is best read as a medium-horizon economic-link signal. Two caveats matter when interpreting the results.

Static graph. The released graph is built from FY2024 filings and applied to returns from 2015 onward. Returns enter only through lagged peer returns, so there is no price leakage, but for earlier months the descriptions may reflect relationships that did not yet exist. The monthly results therefore measure the historical return-relevance of the released links. The filing-date test in Section 5.1 shows that a graph built from FY2023 descriptions remains predictive after those filings were available; a fully point-in-time evaluation would require rebuilding the graph at each filing date. Relatedly, the universe consists of recent index constituents,

so the benchmark is tilted toward large US and East Asian companies.

Statistical strength. Under Newey–West inference and false-discovery control, TW-to-TW is the one market pair that remains individually significant, and the LLM-filter gains are positive for every model tested with confidence intervals that include zero once events are clustered by source stock. We read the 5×5 matrix as exploratory and RQ3 as robustness evidence. The II-GNN and CSPO baselines are controlled implementations of those methods' relation mechanisms on CrossAlpha inputs.

Ethics Statement

CrossAlpha is built from public corporate filings released through SEC EDGAR, EDINET, TWSE/MOPS, DART/KRX, and HKEX. We add no personal data, individual identities, or non-public material information, and release standardised category text rather than redistributing raw filings. All return, IC/ICIR, Sharpe, and drawdown numbers are historical academic measurements of signal quality and do not constitute investment advice, guarantees of future performance, or recommendations to trade any security. Users who deploy CrossAlpha-derived signals in live trading are responsible for their own regulatory, compliance, and risk-management obligations.

Reproducibility Statement

The release includes the artefacts needed to reproduce the reported tables and figures: standardised ten-category text for $\sim$10,700 firm-years, PCA-whitened embeddings, the $\sim$19M-edge semantic graph, aligned daily OHLCV data, source-stock and event lists, the Distiller, Linker, and event-tracing prompts, the frozen LLM-filter labels for all 606 events, robustness controls, and JSON outputs for the reported metrics. All reported evaluations replay from these frozen artefacts without any model API call. The monthly and daily backtesting harnesses are CPU-bound and run on a single workstation. Model snapshots, sampling settings, and refresh cost are reported in Appendix E to make future dataset updates auditable.

Acknowledgements

The authors gratefully acknowledge the support of MS Capital for this research. Zhaomin Wu is partially supported by a National Research Foundation (NRF) Postdoctoral Award. This research is supported by the National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and Infocomm Media Development Authority.

References

Daron Acemoglu, Vasco M. Carvalho, Asuman Ozdaglar, and Alireza Tahbaz-Salehi. 2012. The network origins of aggregate fluctuations. *Econometrica*, 80(5):1977–2016.

Dogu Araci. 2019. FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Jean-Noël Barrot and Julien Sauvagnat. 2016. Input specificity and the propagation of idiosyncratic shocks in production networks. *The Quarterly Journal of Economics*, 131(3):1543–1592.

Geert Bekaert, Robert J. Hodrick, and Xiaoyan Zhang. 2009. International stock return comovements. *The Journal of Finance*, 64(6):2591–2626.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300.

Laura Bini, Francesco Dainelli, and Francesco Giunta. 2016. Business model disclosure in the strategic report: Entangling intellectual capital in value creation process. *Journal of Intellectual Capital*, 17(1):83–102.

Nicholas Bloom, Mark Schankerman, and John Van Reenen. 2013. Identifying technology spillovers and product market rivalry. *Econometrica*, 81(4):1347–1393.

Christian Breitung and Sebastian Mueller. 2025. Global business networks. *Journal of Financial Economics*, 166:104007.

Maryam Bukhari, Muazzam Maqsood, and Asma Sattar. 2025. A novel inter-intra graph neural networks for stock price forecasting modeling cross-border relationships. *Expert Systems with Applications*, 286:127907.

Leland Bybee, Bryan T. Kelly, Asaf Manela, and Dacheng Xiu. 2024. Business news and business cycles. *The Journal of Finance*, 79(5):3443–3492.

Andy Chung and Kumiko Tanaka-Ishii. 2023. Modeling momentum spillover with economic links discovered from financial documents. In *Proceedings of the 4th ACM International Conference on AI in Finance*, pages 490–497. ACM.

Lauren Cohen and Andrea Frazzini. 2008. Economic links and predictable returns. *The Journal of Finance*, 63(4):1977–2011.

Yang Du, Biao Li, Zhichen Lu, and Gang Kou. 2025. Modeling hybrid firm relationships with graph neural networks for stock investment decisions. *Decision Support Systems*, 198:114528.

Cheol S. Eun and Sangdal Shim. 1989. International transmission of stock market movements. *Journal of Financial and Quantitative Analysis*, 24(2):241–256.

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

Georgios Fatouros, Konstantinos Metaxas, John Soldatos, and Dimosthenis Kyriazis. 2025. MarketSenseAI 2.0: Enhancing stock analysis through LLM agents. *arXiv preprint arXiv:2502.00415*.

Fuli Feng, Xiangnan He, Xiang Wang, Cheng Luo, Yiqun Liu, and Tat-Seng Chua. 2019. Temporal relational ranking for stock prediction. *ACM Transactions on Information Systems*, 37(2):27:1–27:30.

Kristin J. Forbes and Menzie D. Chinn. 2004. A decomposition of global linkages in financial markets over time. *The Review of Economics and Statistics*, 86(3):705–722.

Diego García and Øyvind Norli. 2012. Geographic dispersion and stock returns. *Journal of Financial Economics*, 106(3):547–565.

Richard C. Grinold and Ronald N. Kahn. 2000. *Active Portfolio Management: A Quantitative Approach for Producing Superior Returns and Controlling Risk*, 2nd edition. McGraw-Hill.

Yasushi Hamao, Ronald W. Masulis, and Victor Ng. 1990. Correlations in price changes and volatility across international stock markets. *The Review of Financial Studies*, 3(2):281–307.

Campbell R. Harvey, Yan Liu, and Heqing Zhu. 2016. . . . and the cross-section of expected returns. *The Review of Financial Studies*, 29(1):5–68.

Bernard Herskovic. 2018. Networks in production: Asset pricing implications. *The Journal of Finance*, 73(4):1785–1818.

Gerard Hoberg and Gordon Phillips. 2010. Product market synergies and competition in mergers and acquisitions: A text-based analysis. *The Review of Financial Studies*, 23(10):3773–3811.

Gerard Hoberg and Gordon Phillips. 2016. Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, 124(5):1423–1465.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Hong Kong Exchanges and Clearing. 2025. Monthly market highlights. https://www.hkex.com.hk/Market-Data/Statistics/Consolidated-Reports/Monthly-Market-Highlights.

Yikuan Huang, Zheqi Fan, Kaiqi Hu, and Yifan Ye. 2026. Cross-stock predictability via LLM-augmented semantic networks. *arXiv preprint arXiv:2604.19476*.

Timotej Jagrič and Aljaž Herman. 2026. Ai-driven peer company identification: A semantic text-similarity approach beyond traditional industry classification systems. *Business Systems Research: International journal of the Society for Advancing Innovation and Research in Economy*, 17(1):204–222.

Japan Exchange Group. 2025. JPX report 2025. https://www.jpx.co.jp/english/corporate/investor-relations/ir-library/.

Zuben Jin. 2024. Business aspects in focus, investor underreaction and return predictability. *Journal of Corporate Finance*, 84:102525.

Rasmus Jørgensen, Oliver Brandt, Mareike Hartmann, Xiang Dai, Christian Igel, and Desmond Elliott. 2023. MultiFin: A dataset for multilingual financial NLP. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 894–909, Dubrovnik, Croatia. Association for Computational Linguistics.

Agnan Kessy, Alex Lewin, and Korbinian Strimmer. 2018. Optimal whitening and decorrelation. *The American Statistician*, 72(4):309–314.

Alex G. Kim, Maximilian Muhn, and Valeri V. Nikolaev. 2024. Financial statement analysis with large language models. *arXiv preprint arXiv:2407.17866*.

Raehyun Kim, Chan Ho So, Minbyul Jeong, Sanghoon Lee, Jinkyu Kim, and Jaewoo Kang. 2019. HATS: A hierarchical graph attention network for stock movement prediction. *arXiv preprint arXiv:1908.07999*.

Mervyn A. King and Sushil Wadhwani. 1990. Transmission of volatility between stock markets. *The Review of Financial Studies*, 3(1):5–33.

Sida Lin, Yankai Chen, Yiyan Qi, Chenhao Ma, Bokai Cao, Yifei Zhang, Xue Liu, and Jian Guo. 2025. CSPO: Cross-market synergistic stock price movement forecasting with pseudo-volatility optimization. In *Companion Proceedings of the ACM Web Conference 2025*, pages 354–363. ACM.

Jiaxin Liu, Yi Yang, and Kar Yan Tam. 2024. Beyond surface similarity: Detecting subtle semantic shifts in financial narratives. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2641–2652, Mexico City, Mexico. Association for Computational Linguistics.

Jing Liu, Maria Grith, Xiaowen Dong, and Mihai Cucuringu. 2026. A bipartite graph approach to U.S.–China cross-market return forecasting. *arXiv preprint arXiv:2603.10559*.

Alejandro Lopez-Lira and Yuehua Tang. 2023. Can ChatGPT forecast stock price movements? Return predictability and large language models. *arXiv preprint arXiv:2304.07619*.

Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65.

Yueran Ma. 2019. Nonfinancial firms as cross-market arbitrageurs. *The Journal of Finance*, 74(6):3041–3087.

Lior Menzly and Oguzhan Ozbas. 2010. Market segmentation and cross-predictability of returns. *The Journal of Finance*, 65(4):1555–1580.

Volkan Muslu, Suresh Radhakrishnan, K. R. Subramanyam, and Dushyantkumar Lim. 2015. Forward-looking MD&A disclosures and the information environment. *Management Science*, 61(5):931–948.

Whitney K. Newey and Kenneth D. West. 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708.

OpenAI. 2024. New embedding models and API updates. https://openai.com/index/new-embedding-models-and-api-updates/. Accessed: 2024-01-25.

OpenAI. 2025a. GPT-4.1 and GPT-4.1-mini. https://openai.com/index/gpt-4-1/. Accessed: 2025-04-14.

OpenAI. 2025b. GPT-5 model documentation. https://platform.openai.com/docs/models/gpt-5.

Qwen Team. 2025. Qwen3 technical report. arXiv preprint arXiv:2505.09388.

David E. Rapach, Jack K. Strauss, and Guofu Zhou. 2013. International stock return predictability: What is the role of the united states? *The Journal of Finance*, 68(4):1633–1662.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992.

Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Rajiv Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When FLUE meets FLANG: Benchmarks and large pre-trained language model for financial domain. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2335.

William F. Sharpe. 1994. The Sharpe ratio. *Journal of Portfolio Management*, 21(1):49–58.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.

Takehiro Takayanagi, Hiroki Sakaji, and Kiyoshi Izumi. 2024. SETN: Stock embedding enhanced with textual and network information. *arXiv preprint arXiv:2408.02899*.

Yixuan Tang and Yi Yang. 2025. FinMTEB: Finance massive text embedding benchmark. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3620–3638, Suzhou, China. Association for Computational Linguistics.

Paul C. Tetlock. 2007. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168.

Tuomo Vuolteenaho. 2002. What drives firm-level stock returns? *The Journal of Finance*, 57(1):233–264.

Chi-Tai Wang and Chui-Sheng Chiu. 2014. Competitive strategies for taiwan's semiconductor industry in a new world economy. *Technology in Society*, 36:60–73.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788.

World Federation of Exchanges. 2025. Market statistics. https://www.world-exchanges.org/our-work/statistics.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. 2023. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*.

Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. 2024. TradingAgents: Multi-agents LLM financial trading framework. *arXiv preprint arXiv:2412.20138*.

Zhiyu Zhang, Zheng Qiao, Yao Ge, and Zhe Shen. 2026. Uncovering interfirm links through textual topic similarity: A comomentum analysis in financial markets. *The British Accounting Review*, 58(2):101446.

Appendix

Table of Contents

A Data-Source Portals and Usage Policies **15**

B Case Studies: Cross-Market Peer Discovery **15**

B.1 TSMC (2330.TW): Global Semiconductor Peers . . . 15

B.2 Toyota Motor (7203.T): Automotive and Cross-Sector Peers . . . 15

B.3 Apple (AAPL) → Japan: Linker-Inferred Supply Chain . . . 15

B.4 NVIDIA (NVDA) → Japan: Gaming and Testing Links . . . 15

B.5 Boeing (BA) → Japan: Direct Disclosure Relationships . . . 15

B.6 Summary of Case Study Findings . . . 16

C Additional Experimental Results **16**

C.1 Evaluation Metric Definitions . . . 16

C.2 Sector and Size Neutralization . . . 17

C.3 Baseline Peer Definitions . . . 17

C.4 Market Selection for RQ1 . . . 18

C.5 Statistical Inference and Multiplicity . . . 18

C.6 Filing-Date Sensitivity of the Static Graph . . . 18

C.7 Sample Composition, Size, and Liquidity Controls . . . 19

C.8 Controlled Comparison with Prior Cross-Market Methods . . . 19

C.9 Event Strategy Robustness . . . 19

C.10 LLM Extraction Quality, Coverage, and Noise Sensitivity . . . 22

C.11 Dense Graph Construction Details . . . 22

C.12 Hyperparameter Sensitivity . . . 23

C.13 Embedding-Model Provider Robustness . . . 24

C.14 Open-Model Portability . . . 24

D More Related Work **25**

D.1 Financial NLP Benchmarks . . . 25

D.2 Cross-Market and Graph-Based Return Models . . . 25

D.3 Text-Derived Firm Networks and Peer Momentum . . . 26

D.4 LLMs and Agents for Financial Extraction . . . 26

D.5 Supply-Chain Networks in Asset Pricing . . . 26

E Release, Reproducibility, and Construction Cost **27**

E.1 The `AR-Scraper` Toolkit . . . 27

E.2 Construction Cost and Latency . . . 27

F Distiller and Linker Prompts **27**

F.1 Distiller Prompt (ϕ_D): 10-Category Standardisation . . . 27

F.2 Linker Prompt (ϕ_L): Typed Cross-Market Edge Mining . . . 28

F.3 Event-Tracing Prompt (ϕ_L^{event}): Typed Sign on a Hub Event . . . 29

G Grounding the 10-Category Schema Across Regulatory Systems **29**

A Data-Source Portals and Usage Policies

CrossAlpha uses public corporate-disclosure portals for annual reports and a public market-data portal for daily OHLCV prices. Table 8 lists the source portals and the corresponding official usage-policy or terms pages used to check redistribution and access constraints. We release standardised category text rather than raw filings, and users refreshing the dataset should re-check the linked policies because source terms may change.

B Case Studies: Cross-Market Peer Discovery

This appendix reports five examples used for manual inspection of the released cross-market graph. The examples are drawn from the Japan TOPIX 500 universe; peer rankings are based on cosine similarity of LLM-standardized business description embeddings (`text-embedding-3-large`) computed over fiscal year 2022–2024 filings. They are descriptive checks of the constructed links and are not used in the return-prediction evaluation.

B.1 TSMC (2330.TW): Global Semiconductor Peers

Taiwan Semiconductor Manufacturing Company (TSMC, TSE:2330) is the world's largest dedicated semiconductor foundry. Table 9 reports the top-20 cross-market peers identified by CrossAlpha. The nearest-neighbor list is concentrated in foundry and memory firms: SMIC (SEHK:0981) ranks first, Global Foundries (NASDAQ:GFS) second, and SK Hynix (KRX:000660) third. SUMCO, a silicon wafer supplier, ranks fourth despite operating in an adjacent sub-segment.

The market composition of the top-20 peers is: US (12), KR (4), JP (3), HK (1). All top-10 peers fall within the Technology sector (100% same-sector precision@10), although sector labels are not used as inputs. The prevalence of U.S. names (equipment makers such as AMAT, LRCX, KLAC alongside fabless designers NVDA, TXN) reflects the tight semantic overlap between foundry operations and the semiconductor equipment and design ecosystem as described in annual reports.

B.2 Toyota Motor (7203.T): Automotive and Cross-Sector Peers

Toyota Motor Corporation (TSE:7203) is a useful example for inspecting cross-sector nearest neighbors. Table 10 reports the top-10 peers; the full top-30 set contains 47% same-sector (Automotive) names alongside a set of cross-sector discoveries in technology, retail, and mobility.

The cross-sector names at ranks 5 (Oracle) and 8 (Apple) are consistent with the overlap between automotive and technology annual reports, where vehicle manufacturers increasingly discuss software-defined platforms, in-vehicle operating systems, and data analytics infrastructure. Further in the top-30, mobility-platform companies Uber (UBER), Lyft (LYFT), and Aurora Innovation (AUR) appear, reflecting shared business description language around autonomous driving, fleet management, and on-demand transportation. The 47% same-sector precision@30 indicates that the list is not simply a sector lookup.

B.3 Apple (AAPL) → Japan: Linker-Inferred Supply Chain

The Linker (ϕ_L) parses 10-K filings to extract supplier and customer relationships that are not directly disclosed in structured databases. Table 11 reports Japanese suppliers inferred for Apple from annual report disclosures, including those found in the suppliers' own filings, not solely in Apple's 10-K.

None of these relationships were directly disclosed in Apple's 10-K filings; they are based on Japanese companies' own annual report disclosures referencing Apple as a major customer. This is a common disclosure pattern: suppliers often name major customers, while customers may not list all suppliers. The Linker therefore reads *both sides* of the supply-chain relationship graph.

B.4 NVIDIA (NVDA) → Japan: Gaming and Testing Links

NVIDIA's Japanese peer set combines downstream gaming console customers (Sony, Nintendo) with an upstream testing equipment supplier (Advantest). These relationships appear because Sony and Nintendo disclose GPU vendor relationships in their product description sections, and Advantest references NVIDIA SoC platforms in its equipment disclosures. Advantest appears in both the Apple and NVIDIA peer sets (Tables 11 and 12), giving one example of a shared upstream equipment link.

B.5 Boeing (BA) → Japan: Direct Disclosure Relationships

Unlike the Apple and NVIDIA cases, Boeing's Japanese supplier relationships are directly named

Table 8: Official data-source portals and usage-policy links for CrossAlpha.

Market	Data stream	Source portal	Usage-policy / terms link
US	Annual reports	SEC EDGAR, https://www.sec.gov/edgar	SEC website policies and fair-access guidance, https://www.sec.gov/privacy and https://www.sec.gov/os/accessing-edgar-data
JP	Annual reports	EDINET, https://disclosure2.edinet-fsa.go.jp	FSA EDINET API terms, https://disclosure2dl.edinet-fsa.go.jp/guide/static/disclosure/download/ESE140191.pdf
TW	Annual reports	MOPS, https://mops.twse.com.tw	TWSE website terms, https://wwwc.twse.com.tw/zh/terms/use.html
KR	Annual reports	DART, https://dart.fss.or.kr	OpenDART terms, https://opendart.fss.or.kr/intro/terms.do
HK	Annual reports	HKEXnews, https://www.hkexnews.hk	HKEX website terms and conditions, https://www.hkex.com.hk/Global/Exchange/Terms-of-Use
All markets	Daily prices	Yahoo Finance, https://finance.yahoo.com	Yahoo terms and finance-data restrictions, https://legal.yahoo.com/us/en/yahoo/terms/otos/index.html and https://legal.yahoo.com/us/en/yahoo/terms/product-atos/apiforydn/index.html

in Boeing's own 10-K filings under the supplier diversity and key suppliers disclosures. Mitsubishi Heavy Industries is identified as the manufacturer of wing box structures for the 787 Dreamliner, and Kawasaki Heavy Industries is identified as responsible for fuselage sections and landing gear doors. This example contrasts with the Apple and NVIDIA cases: Boeing discloses these suppliers directly, whereas some consumer-electronics links are observed from supplier-side filings. The released graph therefore contains both directly disclosed links and links inferred from counterpart filings.

B.6 Summary of Case Study Findings

The examples serve as sanity checks for three aspects of the released graph: sector recovery, cross-sector nearest-neighbor retrieval, and supplier-side disclosure links. Evidence for return predictability is evaluated separately in the main experiments.

C Additional Experimental Results

C.1 Evaluation Metric Definitions

For the portfolio metrics, let r_t^{LS} be the net monthly long–short return after a 2 bp one-way cost and $V_t = \prod_{\tau=1}^{t}(1 + r_\tau^{\mathrm{LS}})$ the corresponding wealth curve. **(1)** *IC* is the monthly rank correlation between factor values and forward returns,

$$\mathrm{IC}_t = \mathrm{corr}_{\mathrm{rank}}\Big(f_i(t), r_i^{\mathrm{fwd}}(t)\Big); \tag{4}$$

(2) *ICIR* is its annualised information ratio,

$$\mathrm{ICIR} = \frac{\frac{1}{T}\sum_{t=1}^{T}\mathrm{IC}_t}{\hat{\sigma}_{\mathrm{IC}}}\sqrt{12}; \tag{5}$$

(3) *Sharpe* is the annualised long–short Sharpe ratio,

$$\mathrm{Sharpe} = \frac{\overline{r^{\mathrm{LS}}}}{\hat{\sigma}(r^{\mathrm{LS}})}\sqrt{12}; \tag{6}$$

(4) *MaxDD* is the maximum drawdown of the wealth curve,

$$\mathrm{MaxDD} = \min_t\left(\frac{V_t}{\max_{\tau\le t} V_\tau} - 1\right); \tag{7}$$

(5) *Ret* is the annualised long–short return after cost,

$$\mathrm{Ret} = 12\,\overline{r^{\mathrm{LS}}}. \tag{8}$$

Finally, **(6)** *CumRet* is the ending cumulative long–short return over the evaluation window,

$$\mathrm{CumRet} = \sum_{t=1}^{T} r_t^{\mathrm{LS}}. \tag{9}$$

Table 9: Top-20 cross-market peers of TSMC (2330.TW) by embedding similarity.

Rank	Ticker	Company	Market	Sector
1	0981.HK	SMIC	HK	Technology
2	GFS	GlobalFoundries	US	Technology
3	000660.KS	SK Hynix	KR	Technology
4	3436.T	SUMCO	JP	Technology
5	INTC	Intel	US	Technology
6	TXN	Texas Instruments	US	Technology
7	AMAT	Applied Materials	US	Technology
8	LRCX	Lam Research	US	Technology
9	KLAC	KLA Corporation	US	Technology
10	005930.KS	Samsung Electronics	KR	Technology
11	ASML	ASML Holding	US	Technology
12	MU	Micron Technology	US	Technology
13	6723.T	Renesas Electronics	JP	Technology
14	NVDA	NVIDIA	US	Technology
15	6954.T	FANUC	JP	Technology
16	ON	ON Semiconductor	US	Technology
17	STM	STMicroelectronics	US	Technology
18	8035.T	Tokyo Electron	JP	Technology
19	009150.KS	Samsung SDI	KR	Technology
20	ADI	Analog Devices	US	Technology

C.2 Sector and Size Neutralization

For the cross-market source comparisons we neutralize each monthly cross-section in two steps. First, we demean the factor within GICS sector, removing each sector's average factor level on that date. Second, we standardize the sector-demeaned factor and regress it on standardized log market capitalisation, $\ln(\text{close} \times \text{shares outstanding})$, keeping the cross-sectional residual. The resulting factor is orthogonal to both sector membership and firm size, and the neutralized rank ICIR (Equation (5)) is computed between this residual and forward returns. A stricter variant used for robustness additionally regresses out industry- and own-momentum and winsorizes the cross-section at the 1% level before standardization.

C.3 Baseline Peer Definitions

All RQ1 baselines share the monthly peer-momentum pipeline of Section 4: the same 12-month sector-relative source-market returns, the same sigmoid top-1% weighting (Equation (10)), portfolio sort, and neutralization. They differ only in how the peer weight α_{ij} between a target stock i and a candidate j is set, and—for the domestic baseline—in which market supplies the candidates.

- **CrossAlpha text (cross-market).** α_{ij} is the released US→JP text similarity (cosine of PCA-whitened category embeddings, Equation (2)); candidates are US Russell 1000 stocks.
- **Domestic text (JP→JP).** The identical text pipeline, but candidates are the target's own market (JP TOPIX 500) and peer returns are JP sector-relative, isolating whether cross-market text adds information beyond within-market text.
- **GICS sector.** Each target inherits the equal-weighted mean sector-relative return of the US peers sharing its GICS sector (no text and no similarity weighting); this is the industry-classification peer set.
- **Return correlation (252d).** α_{ij} is the Pearson correlation of trailing 252-day daily returns between i and US candidate j (minimum 120 overlapping days), percentile-ranked and passed through the same sigmoid (Equation (10)); this is the standard backward-looking, price-based peer definition.

Table 10: Top-10 cross-market peers of Toyota (7203.T) by embedding similarity.

Rank	Ticker	Company	Market	Sector
1	TSLA	Tesla	US	Consumer Discretionary
2	F	Ford Motor	US	Consumer Discretionary
3	GM	General Motors	US	Consumer Discretionary
4	RIVN	Rivian Automotive	US	Consumer Discretionary
5	ORCL	Oracle	US	Technology
6	AN	AutoNation	US	Consumer Discretionary
7	KMX	CarMax	US	Consumer Discretionary
8	AAPL	Apple	US	Technology
9	HOG	Harley-Davidson	US	Consumer Discretionary
10	CVNA	Carvana	US	Consumer Discretionary

Table 11: Linker-inferred Japanese supply-chain peers for Apple (AAPL).

Ticker	Company	Relationship Type	Component / Role
6758.T	Sony Group	Supplier (inferred)	Image sensors (camera modules)
6981.T	Murata Mfg.	Supplier (inferred)	Multilayer ceramic capacitors (MLCCs)
6762.T	TDK	Supplier (inferred)	Lithium-ion battery cells
6857.T	Advantest	Supplier (inferred)	Chip test and measurement equipment

Because all four aggregate peer returns identically, any performance gap reflects the peer definition alone.

C.4 Market Selection for RQ1

RQ1 uses US-to-Japan as the controlled peer-definition comparison because the United States provides a broad, globally connected source-peer universe, and Japan provides a large, fully covered target universe with the lowest single-sector concentration among the Asian markets (Table 15). RQ2 then applies the same construction to all ordered market pairs, including the more sector-concentrated Taiwanese and Korean targets, to test whether the source-market advantage generalises across sector compositions.

C.5 Statistical Inference and Multiplicity

For RQ1 and RQ2 we apply Newey–West inference with Bartlett lag 3 (Newey and West, 1987) to the exact monthly IC series underlying the reported tables, preserving their universes, horizons, lag, and neutralization. Because the Information-Geography matrix tests 25 market pairs simultaneously, we apply the Benjamini–Hochberg false-discovery-rate procedure at q=0.10 (Benjamini and Hochberg, 1995), which bounds the expected proportion of false positives among the cells declared significant. For RQ3 we pair graph-only and filtered per-event returns on the same 606 events across 25 source stocks at K=20, cluster standard errors by source stock, and apply Holm correction across basket sizes (Holm, 1979). Table 16 summarises the results. Only the strongest RQ2 cell survives FDR control; the US-to-Japan RQ1 result and the RQ3 filter increments are positive but not individually significant, and the corresponding claims in the main text are stated accordingly.

C.6 Filing-Date Sensitivity of the Static Graph

The released graph is built from FY2024 descriptions and applied to January 2015–December 2025 returns, so pre-FY2024 months contain source-side structural look-ahead. To quantify it, we run the final-neutralized monthly US-to-Japan task (Russell 1000 source peers, TOPIX 500 targets, 12-month lookback, one-month lag) and replace only the US-side descriptions with ones distilled from earlier filings, holding the Japanese descriptions and the evaluation protocol fixed (Table 17). Over the common July 2024–December 2025 window, ICIR declines from 0.57 with the released descriptions to 0.37 with FY2023 descriptions and 0.20 with FY2022 descriptions, quantifying attenuation

Table 12: Linker-inferred Japanese supply-chain peers for NVIDIA (NVDA).

Ticker	Company	Relationship Type	Component / Role
6758.T	Sony Group	Customer (inferred)	PlayStation 5 GPU (Oberon SoC)
7974.T	Nintendo	Customer (inferred)	Nintendo Switch SoC (Tegra X1)
6857.T	Advantest	Supplier (inferred)	GPU test and verification equipment

Table 13: Japanese supply-chain peers for Boeing (BA): direct 10-K disclosures.

Ticker	Company	Disclosure Source	Component / Role
7011.T	Mitsubishi Heavy Industries	Boeing 10-K (direct)	Wing box structures (787 Dreamliner)
7012.T	Kawasaki Heavy Industries	Boeing 10-K (direct)	Fuselage sections, landing gear doors

as the source descriptions become older. The decay is not uniform: after restricting the released graph to the same US peer set as the FY2023 graph, FY2023 descriptions reach 0.37 versus 0.32 for the released descriptions. Most importantly, the graph constructed from FY2023 US descriptions remains predictive for returns from July 2024 onward, after those source descriptions were available, so future US-side text does not account for the central finding. A fully point-in-time evaluation would require reconstructing both sides of the graph at each filing date.

C.7 Sample Composition, Size, and Liquidity Controls

The covered firms are large and liquid, so we test whether the US-to-Japan signal could arise from broad large-cap, sector, or liquidity co-movement rather than from the text-derived firm links. All variants use the final-neutralized monthly US-to-Japan task (GICS-sector demeaning, log-market-cap residualisation, 12-month lookback, one-month lag). Each month, Russell 1000 peers are sorted into size and liquidity thirds using market capitalisation (close $\times$ shares outstanding) and median daily dollar volume (close $\times$ volume). The *matched-random* control permutes US peer identities only within the same GICS-sector $\times$ size-tercile $\times$ liquidity-tercile cell, preserving composition while removing the identity of the linked firm. The *liquidity-exclusion* control removes the most-liquid US decile, re-formed monthly (90.6 firms on average, 11.9% of graph weight). For the target-side subgroup analysis, eligible Japanese firms are sorted each month into three approximately equal-sized groups by the same measures. Table 18 reports the results: the matched-random ICIR collapses around zero and no seed reaches the CrossAlpha value, while excluding the most-liquid US names leaves the result essentially unchanged. The signal is strongest among high-market-cap Japanese targets, and the low-liquidity group remains positive.

C.8 Controlled Comparison with Prior Cross-Market Methods

To test whether the benchmark discriminates disclosure-derived relations from price-derived cross-market mechanisms, we implement the cross-market relation mechanism of II-GNN (Bukhari et al., 2025) and CSPO (Lin et al., 2025) using only CrossAlpha-available inputs, and evaluate them under the final-neutralized monthly US-to-Japan protocol (Russell 1000 to TOPIX 500, 101 common months, 12-month lookback, one-month lag, identical labels, neutralization, and costs). The original methods rely on additional modalities that are outside the benchmark (Table 19), so these are *controlled, benchmark-compatible implementations* of each method's relation mechanism. Table 20 shows that CrossAlpha outperforms every controlled baseline in IC, signed ICIR, and net long–short Sharpe; the two smaller paired IC p values are suggestive after Holm correction (both 0.052).

C.9 Event Strategy Robustness

Event execution protocol. Source stocks are fixed before the test window: the top-5 firms by 2023-12-31 market capitalisation in each of the five markets, restricted to firms with a 10-K-equivalent filing, yielding 25 stocks (Appendix C.9). On the same event set, *Graph alone* equal-weights the source stock's top-K neighbours, with $K \in \{10, 20, 30, 60\}$, pooled across all five markets.

Table 14: Summary statistics across the five case studies.

Company	Peer Set Size	Same-Sector %	Key Finding
TSMC (2330.TW)	Top 20	100% (top-10)	US/KR/JP foundry ecosystem
Toyota (7203.T)	Top 30	47%	Oracle, Apple, Uber as peers
Apple (AAPL)	4 JP links	Linker-inferred	Sony, Murata, TDK, Advantest
NVIDIA (NVDA)	3 JP links	Mixed	Sony PS5, Nintendo Switch, Advantest
Boeing (BA)	2 JP links	Direct 10-K	MHI wing box, KHI fuselage

Table 15: Target-market characteristics used to select the RQ1 pair. Market capitalisation is the 2025 exchange-level figure (Japan Exchange Group, 2025; World Federation of Exchanges, 2025; Hong Kong Exchanges and Clearing, 2025); coverage counts firms with all ten Distiller fields populated.

Target	2025 mkt. cap	Ten-field coverage	Largest sector weight
Japan	US$6.3T	453/453	Industrials, 22.2%
Taiwan	US$2.97T	1,040/1,040	Technology, 73.4%
South Korea	US$2.69T	612/612	Technology, 57.6%
Hong Kong	≈US$6.1T	602/602	Financial Services, 33.6%

Graph + LLM-filter-labelled co-movers uses the same graph ranking but keeps only candidates that the GPT-5 LLM filter marks as confident co-movers before forming the K-name basket. The LLM filter never proposes firms outside the graph shortlist. Execution is strict $t{+}2$: buy at the $D{+}2$ open and sell at the $D{+}3$ open. We leave $D{+}1$ as a no-action buffer because different markets open in different time zones, so a $D{+}1$ trade may be infeasible for some source–target pairs, a standard alignment concern in cross-market return tests (Bekaert et al., 2009; Rapach et al., 2013). All returns are market-relative.

Table 7 (main text) reports return and Sharpe metrics across K. Table 21 adds the corresponding per-event t-statistics on the same 606 GPT-5 post-cutoff events.

Matched-neighbour null. To test whether the graph result is simply a by-product of market allocation or sector composition, Table 22 keeps the same 25 source stocks, 606 positive events, strict $t+2$ execution, and market-relative returns, but replaces each graph basket with 200 fixed-seed random baskets. The first null matches the graph basket's market allocation; the second matches each graph neighbour's market and GICS sector where available, falling back to the same market if a sector bucket is too sparse. The graph-selected neighbours sit at or above the 95th percentile of both nulls at every $K \in \{10, 20, 30, 60\}$ (graph-percentile vs. market null: 100 / 99 / 100 / 100; vs. market+sector null: 100 / 95 / 99.5 / 99), indicating that the event spillover is tied to the text-derived neighbour ranking rather than to coarse market or sector exposure.

Source-stock-cluster uncertainty. Because events from the same source stock are not independent, we also bootstrap by resampling source stocks rather than individual events. Over 2,000 source-stock-cluster bootstrap draws, the graph-alone $K = 30$ book has a per-event mean 95% interval of $[-0.018\%, +0.154\%]$ and $P(\bar{r} > 0) = 0.942$. This check treats the source stock as the dependence unit and is supportive but deliberately conservative; the interval crosses zero, so we use it as a robustness diagnostic rather than as the main evidence.

LLM-filter model controls and replay specification. The main RQ3 filter is GPT-5, requested through the alias `gpt-5` whose pinned catalog snapshot is `gpt-5-2025-08-07`; temperature is not set in the request (API default). The artifact freezes the request identifier, prompt, shortlist, per-firm labels for all 25 source stocks, event list, and outputs, so every reported RQ3 number replays offline without an API call. All 606 evaluated events occur after the documented 2024-09-30 knowledge cutoff (OpenAI, 2025b), and the filter receives only firm descriptions and the frozen graph shortlist, never event outcomes. To isolate what the LLM adds, Table 23 applies the same prompt and keep rule to GPT-5.5 and to the open-weight Qwen3-235B-A22B-Instruct-2507 (Qwen Team, 2025) (temperature 0), and adds a no-LLM same-sector rule, while holding fixed the same 606 events, 25 source stocks, graph shortlists, and return protocol at $K{=}20$ (a cap, because filters may retain fewer firms). The open-weight model reproduces the direction and most of the magnitude (Sharpe 1.40 versus 1.50), while the rule-based filter does not improve on the graph. As

Table 16: Multiplicity-aware inference for the headline results. NW: Newey–West with Bartlett lag 3. CI: 95% confidence interval for the mean monthly IC. RQ3 increments are mean per-event return gains over the graph-alone basket at $K=20$.

Result	Statistic	IC / increment	Signed ICIR	Uncertainty	Multiplicity outcome
US-to-Japan RQ1	mean IC	0.0094	0.39	NW $t=1.55$; CI $[-0.0026, 0.0213]$	Not individually significant
Strongest 5×5 cell (TW-to-TW)	mean IC	–	0.87	NW $t=3.04$	Survives BH-FDR at $q=0.10$
Full 5×5 matrix	25 IC tests	–	–	–	1/25 cells survives BH-FDR at $q=0.10$
RQ3 filter: GPT-5	per-event increment	$+0.057$ pp	–	source-cluster $p=0.167$	Holm-adjusted $p=0.501$
RQ3 filter: GPT-5.5	per-event increment	$+0.073$ pp	–	source-cluster $p=0.098$	Holm-adjusted $p=0.294$
RQ3 filter: Qwen3-235B	per-event increment	$+0.042$ pp	–	source-cluster $p=0.154$	Holm-adjusted $p=0.461$

Table 17: Filing-date sensitivity in the final-neutralized monthly US-to-Japan task (signed ICIR). Only the US-side descriptions change.

US-side descriptions	Jan. 2015–Dec. 2025	Jul. 2024–Dec. 2025
Derived from FY2022 filings	0.58	0.20
Derived from FY2023 filings	0.27	0.37
Released (FY2024)	0.52	0.57
Released, restricted to the FY2023 US peer set	0.25	0.32

Table 18: Sample-composition controls in the final-neutralized monthly US-to-Japan task (signed ICIR).

Robustness check	Signed ICIR
CrossAlpha disclosure graph	0.526
Exclude most-liquid US decile, re-formed monthly	0.511
GICS sector × size × liquidity matched random, 10 seeds	-0.027 ± 0.279 (0/10 reach 0.526)
Japanese market-cap terciles, low / middle / high	-0.164 / -0.008 / 0.703
Japanese liquidity terciles, low / middle / high	0.177 / -0.063 / 0.526

Appendix C.5 reports, the per-event increments are positive for every model but not statistically significant after source-stock clustering and Holm correction.

Event case study. To make the RQ3 basket concrete, consider one post-cutoff event from the evaluation window: SK Hynix (000660.KS) on 2025-10-02. The source stock had a market-relative return of $+9.11\%$, consistent with a positive AI-memory / HBM demand shock. CrossAlpha's graph retrieved a global neighbour shortlist containing memory-sector peers and hardware customers; the LLM filter then labelled confident positive co-movers such as Taiwanese semiconductor peers 2408.TW, 6531.TW, 8271.TW, and 2344.TW, plus hardware-exposed 4967.TW. Under the same strict $t+2$ rule as the main experiment, the confident co-mover basket with available cross-market returns earned an average market-relative return of $+4.53\%$ over the one-day $D+2$ open-to-$D+3$ open holding window, with 15 of 17 names positive. This example is descriptive rather than a separate test: the reported RQ3 results pool all 606 post-cutoff events.

The LLM filter also removes graph neighbours that are textually close but have the wrong exposure for the event. For an NVIDIA AI-demand shock, the graph shortlist contains Nintendo (7974.T), a plausible textual neighbour because both firms share gaming and GPU-related business language: NVIDIA historically sells GPUs into gaming devices, while Nintendo discloses console hardware, game software, and platform ecosystems. The textual graph therefore retrieves Nintendo as a reasonable broad business neighbour. The LLM filter labels Nintendo as *unrelated* with sign 0, because the event is a data-center AI GPU demand surprise rather than a console-cycle or gaming-demand shock. In that event window, firms tied to AI accelerators, HBM memory, servers, and semiconductor equipment should be more direct co-movers, while Nintendo's economics are driven mainly by Switch hardware, first-party software, and consumer gaming demand. The realised return is consistent with this filter: under the same strict $t+2$ open-to-open rule after NVIDIA's 2024-02-22 AI-demand event, Nintendo's one-day market-relative return was -2.83% (raw return -2.65%), so it did not participate in the positive AI-semiconductor spillover. This is the intended role of the LLM filter: keep firms with direct exposure to the shock and remove semantically nearby firms whose economics should not co-move in that event window.

Source-stock list. The 25 source stocks used in RQ3, by region and Dec-2023 mcap rank: **US Russell 1000**: AAPL, MSFT, GOOG, AMZN, NVDA; **JP TOPIX 500**: 7203.T (Toyota), 8306.T (Mitsubishi UFJ), 6758.T (Sony), 6861.T (Keyence), 8035.T (Tokyo Electron); **TW TWSE**: 2330.TW (TSMC), 2317.TW (Hon Hai / Foxconn), 2454.TW (MediaTek), 2308.TW (Delta Electronics), 2382.TW (Quanta); **KR KOSPI**: 005930.KS (Samsung Electronics), 000660.KS (SK Hynix), 373220.KS (LG Energy Solution), 207940.KS (Samsung Biologics), 005380.KS (Hyundai Motor); **HK Main**: 0700.HK (Tencent), 9988.HK (Alibaba), 1398.HK (ICBC), 0939.HK (CCB), 0005.HK (HSBC).

Table 19: Scope of the controlled prior-method implementations.

Prior method	Original method-specific inputs and architecture	Benchmark-compatible implementation
II-GNN / MEIG (Bukhari et al., 2025)	42.7M event-sentiment tweets, macroeconomic features, and a learned CGAT/GNN over multiple national markets	Inter/intra price-correlation graphs with a strictly walk-forward readout
CSPO (Lin et al., 2025)	An auxiliary futures market, BDP-Former cross-market attention, and a learned pseudo-volatility estimator	US equities as the auxiliary market, cross-market synergy, and pseudo-volatility-weighted walk-forward estimation

Table 20: Controlled prior-method comparison in the final-neutralized monthly US-to-Japan task over 101 common months. Paired p values compare monthly IC against CrossAlpha; Holm correction is across the three comparisons.

Method	Mean IC	Signed ICIR	L/S Sharpe	Paired IC p	Holm-adjusted p
CrossAlpha disclosure graph	0.0110	0.550	0.168	–	–
II-GNN inter/intra aggregation core	0.0021	0.050	0.042	0.523	0.523
II-GNN core + walk-forward ridge readout	−0.0151	−0.498	−0.229	0.023	0.052
CSPO synergy + pseudo-volatility WLS core	−0.0171	−0.552	−0.449	0.017	0.052

Table 21: Per-event t-statistics for the event strategy in Table 7. Same 606 GPT-5 post-cutoff events / 25 source stocks / 1.51 yrs OOS; strict $t+2$ open-to-open, market-relative returns.

Strategy	K	t
Graph alone (CrossAlpha emb. top-K)	10	+1.62
Graph + GPT-5 LLM filter	**10**	**+2.26**
Graph alone	20	+1.00
Graph + GPT-5 LLM filter	**20**	**+1.84**
Graph alone	30	+1.67
Graph + GPT-5 LLM filter	**30**	**+1.78**
Graph alone	60	+1.74
Graph + GPT-5 LLM filter	**60**	**+1.83**

C.10 LLM Extraction Quality, Coverage, and Noise Sensitivity

Human audit. We assess LLM extraction quality via a human-annotated sample of 200 randomly selected filings (25 per market), with three categories randomly chosen per filing, yielding 600 field-level judgments. Annotators rate each field on a 3-point scale: "accurate" (89%), "partial" (9%), and "incorrect" (2%). A *partial* case is a supported but incomplete summary; for example, when a filing describes the firm's primary business in its business section and a smaller segment or regional operation elsewhere, the summary correctly identifies the primary business but omits the secondary segment or region. An *incorrect* case contains a material unsupported statement or omission. The lowest accuracy is for the Korean-language filings (82% accurate, 10% partial, 8% incorrect), consistent with lower LLM pretraining coverage for non-English financial disclosures. Bilingual Hong Kong filings provide a consistency check: the average cosine similarity between English and Chinese sections of the same filing is 0.84 ($\pm$0.06), suggesting that the Distiller produces stable business descriptions when the same issuer discloses in both versions. Construction is fully automated; human supervision is used for stratified validation when the benchmark is extended to a new market, prompt, or model backend.

Coverage audit. A possible bias channel is that firms with complex or less standardised disclosures fail extraction and drop out of the graph. Table 24 shows that this does not occur: every selected firm in all five markets enters the graph, all populated fields exceed 50 characters, and only eight US firms lack one field (`revenue_model`).

Extraction-noise stress test. To test whether extraction errors at the observed worst-language rate could drive the return result, we use the 18% Korean partial-or-incorrect share as the stress level. For each of 10 seeds we reassign 18% of firm-category fields to categories from other firms, rebuild the graph, and rerun the final-neutralized monthly US-to-Korea evaluation with the same 12-month lookback and one-month lag. Wrong-firm reassignment is more severe than incomplete compression, yet the perturbed graph retains 92.7% of the clean signed ICIR (0.260 versus 0.280) and remains positive in all 10 seeds (range 0.116–0.505; Table 25).

C.11 Dense Graph Construction Details

This appendix gives the construction detail condensed in Section 3.3, where the directed cross-market text-similarity graph $\mathcal{G}_{\text{text}} = (V, E_{\text{text}})$ and the firm–firm similarity $s_{\text{text}}(i, j)$ are defined (Equations (1) and (2)). Each node additionally carries a home-market type via $\tau(v) : V \to \mathcal{M}$. We use equal category weights in s_{text} to avoid

Table 22: Matched-neighbour null for the graph-alone event basket. Same 25 source stocks / 606 GPT-5 post-cutoff events / 1.51 yrs OOS / strict $t+2$ open-to-open / market-relative returns. Null entries report the mean per-event return across 200 fixed-seed random baskets and the 95th percentile of the seed-level null distribution.

K	Graph per-ev.	Graph t	Market null mean	Market null p95	Market+sector mean	Market+sector p95
10	$+0.107\%$	$+1.62$	-0.000%	$+0.058\%$	-0.005%	$+0.058\%$
20	$+0.051\%$	$+1.00$	-0.000%	$+0.040\%$	$+0.008\%$	$+0.051\%$
30	$+0.073\%$	$+1.67$	-0.003%	$+0.032\%$	$+0.014\%$	$+0.053\%$
60	$+0.052\%$	$+1.74$	$+0.000\%$	$+0.024\%$	$+0.015\%$	$+0.040\%$

Table 23: Filter model controls over the same graph shortlist at $K=20$ (descriptive Sharpe) and the corresponding reproducibility artifacts.

Filter	Sharpe	Model setting / artifact
Graph alone	0.81	–
GPT-5	1.50	`gpt-5-2025-08-07`; API-default sampling; frozen prompt, shortlist, and 25 label files
GPT-5.5	1.57	same prompt and keep rule; frozen labels
Qwen3-235B (open weights)	1.40	`qwen3-235b-a22b-instruct-2507`; temperature 0; frozen prompt, labels, outputs
Same-sector rule (no LLM)	0.38	keep shortlisted firms sharing the source stock's GICS sector

Table 24: Distillation coverage audit by market.

Market	Parsed / graph firms	All ten fields	Fields $<$ 50 chars
US	880/880	872/880	0
Japan	453/453	453/453	0
Taiwan	1,040/1,040	1,040/1,040	0
Korea	612/612	612/612	0
Hong Kong	602/602	602/602	0

injecting task-specific beliefs about which disclosure fields should matter most; category-specific and learned-weight variants are treated as ablations rather than the default construction.

Peer-selection weights. To form discrete edges $(i, j) \in E_{\text{text}}$ and concentrate signal on the most relevant peers, we apply percentile ranking and a sigmoid transformation:

$$\alpha_{ij}^{\text{text}} = [1 + \exp(-\kappa(\text{rank}_{ij} - \tau))]^{-1}, \quad \kappa = 50, \qquad \tau = 0.99. \tag{10}$$

This assigns near-zero weight to the majority and concentrates $\alpha_{ij}^{\text{text}}$ on the top $\sim$1% most similar peers, roughly 10 stocks per target.

Embedding and whitening. Each standardised category text is embedded using OpenAI `text-embedding-3-large` (OpenAI, 2024), yielding

$$\mathbf{e}_i^{(c)} \in \mathbb{R}^{3072}. \tag{11}$$

Text embeddings are well known to suffer from a long-tailed eigen-spectrum and semantic collapse, where a few dominant directions inflate pairwise cosine similarities (Su et al., 2021); here these directions capture the shared style of the standardised summaries rather than firm-specific content. We address this with PCA (Kessy et al., 2018), projecting embeddings onto an isotropic, lower-dimensional space:

$$\mathbf{z}_i^{(c)} = \sqrt{n} \cdot \mathbf{U}_d \mathbf{\Sigma}_d^{-1} \mathbf{e}_i^{(c)}, \tag{12}$$

where $\mathbf{z}_i^{(c)} \in \mathbb{R}^d$, and $\mathbf{U}_d$ and $\mathbf{\Sigma}_d$ are the top-d components from the SVD of the centred embedding matrix. CrossAlpha releases six dimensionality variants,

$$d \in \{64, 128, 256, 512, 768, 1024\}, \tag{13}$$

so users can revisit the trade-off on new markets.

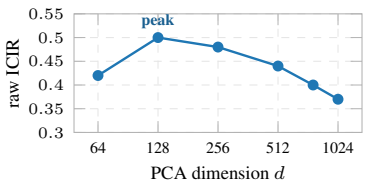
Table 25: Extraction-noise stress test in the final-neutralized monthly US-to-Korea task.

Graph	Signed ICIR	Fraction of clean	Positive seeds	Seed range
Clean	0.280	100%	–	–
18% wrong-firm-category reassignment	0.260	92.7%	10/10	0.116–0.505

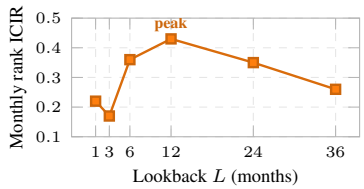
C.12 Hyperparameter Sensitivity

We test how sensitive the cross-market factor is to the two settings that govern the dense-graph pipeline's inductive bias, the PCA-whitening dimension and the peer-return lookback, sweeping each while holding everything else at the defaults used throughout the paper (Figure 8).

Sensitivity around the released defaults. In this sweep, the released defaults are near the best observed settings. Rank ICIR peaks at the PCA-whitening dimension of 128 (0.50) and declines monotonically to 0.37 at 1024, consistent with embedding anisotropy and noisy tail components (Su et al., 2021). The peer-return lookback peaks at 12 months (0.43) and decays on both sides, consistent with the annual-filing cycle and the customer–supplier diffusion timescale (Cohen and Frazzini, 2008; Menzly and Ozbas, 2010). The preferred dimension is mildly market-dependent (e.g.,



(a) Embedding size d.



(b) Lookback L.

Figure 8: Hyperparameter sensitivity of the US-to-JP factor. (a) Raw ICIR peaks at PCA dimension 128 and declines beyond. (b) Monthly rank ICIR peaks at a 12-month lookback and decays on both sides.

US→TW semiconductors favour d=256); all six variants ($d \in \{64, 128, 256, 512, 768, 1024\}$) are released so researchers can test this sensitivity.

C.13 Embedding-Model Provider Robustness

To check whether the text-similarity signal depends on one embedding provider, we run a lightweight robustness check on the US→Japan source–target setting using only the `main_business_segments` field (Table 26). This diagnostic uses the daily peer-momentum harness rather than the headline monthly neutralized pipeline, so we treat it as a representation sanity check rather than a replacement for Table 3. We compare OpenAI `text-embedding-3-large` against two open-source sentence-transformer families (MiniLM/SBERT (Wang et al., 2020; Reimers and Gurevych, 2019) and MPNet/SBERT (Song et al., 2020; Reimers and Gurevych, 2019)), plus FinBERT (Araci, 2019) and a sparse TF-IDF baseline (Salton et al., 1975). All neural and sparse representations produce positive ICIR under the same US-peer / JP-target factor construction; OpenAI remains the strongest single encoder in this diagnostic, while MPNet and TF-IDF are close.

C.14 Open-Model Portability

The released benchmark and all reported evaluations reproduce from frozen artifacts without proprietary model calls. To establish that the construction pipeline also transfers to open backends, we test the encoder and the Distiller separately.

Table 26: Embedding-provider robustness on US→Japan using `main_business_segments` only. ICIR is daily rank-ICIR under the same peer-momentum harness at 5-day and 20-day peer-return lookbacks.

Representation	Dim.	5d ICIR	20d ICIR
OpenAI 3-large	3072	6.33	2.79
MPNet/SBERT	768	6.04	2.73
TF-IDF	5000	6.04	2.74
MiniLM/SBERT	384	5.76	2.58
FinBERT	768	5.38	2.33

Encoder, full universe. Table 27 replaces the OpenAI encoder in the full-universe, final-neutralized monthly US-to-Japan task (Russell 1000 to TOPIX 500, 12-month lookback, one-month lag, same costs) while holding the ten-field text, 128-dimensional whitening, sample, labels, and neutralization fixed. All three open encoders retain positive ICIR; OpenAI remains the strongest in IC and ICIR, while BGE-M3 and MPNet yield higher net long–short Sharpe.

Table 27: Encoder portability in the full-universe final-neutralized monthly US-to-Japan task.

Encoder	IC	Signed ICIR	L/S Sharpe
OpenAI `text-embedding-3-large`	0.01355	0.526	0.147
BGE-M3 (Xiao et al., 2023)	0.00673	0.262	0.313
MPNet-base (Song et al., 2020)	0.00451	0.169	0.271
MiniLM-L6 (Wang et al., 2020)	0.00321	0.118	0.163

Distiller, open-weight diagnostic. On a sector-stratified 300-US × 300-Japan subset, we rerun the Distiller with Qwen3-235B-A22B-Instruct-2507 (Qwen Team, 2025) using the identical prompt at temperature 0 (Table 28). The open model returns valid ten-field JSON for all 600 filings; its category texts have a same-firm/category cosine similarity of 0.93 to the GPT-4.1 texts, and the resulting US-peer rankings agree with the released ones at Spearman 0.75. This is a portability diagnostic on 600 filings; it establishes that the Distiller extends to an open-weight backend.

Table 28: Open-weight Distiller diagnostic on the 300×300 US–Japan subset.

Diagnostic	Result
Valid ten-field JSON	600/600 filings
Same-firm/category cosine vs. GPT-4.1 text	0.93
US-peer ranking agreement (Spearman)	0.75
Hosted inference cost	≈$4

Pretraining contamination. Return outcomes never enter the Distiller prompt, monthly returns are explicitly lagged, and all 606 RQ3 events occur after the documented model cutoff, so the LLM outputs cannot encode the evaluated outcomes.

D More Related Work

This appendix expands the main-paper related-work summary (Section 2) with additional detail on the research areas CrossAlpha draws on.

Cross-market Return Predictability. Classic international-finance work documents return and volatility transmission across equity markets (Eun and Shim, 1989; Hamao et al., 1990; King and Wadhwani, 1990), and later studies show that US returns, international co-movement, and trade links help explain global market linkages (Rapach et al., 2013; Bekaert et al., 2009; Forbes and Chinn, 2004). At the firm level, customer–supplier, same-industry, and production-network links generate predictable return spillovers (Cohen and Frazzini, 2008; Menzly and Ozbas, 2010; Acemoglu et al., 2012; Herskovic, 2018). Recent cross-market forecasting work mainly learns from price/volume series or trading-signal graphs (Lin et al., 2025; Bukhari et al., 2025; Liu et al., 2026), while text-derived firm-link studies either remain mostly domestic or are not released as cross-market return benchmarks (Feng et al., 2019; Kim et al., 2019; Du et al., 2025; Breitung and Mueller, 2025). CrossAlpha fills this gap by releasing annual-report schema text, cross-market semantic links, aligned prices, and factor-evaluation protocols.

Text-based Firm Similarity. Text-based network industries derive peer links from 10-K product descriptions (Hoberg and Phillips, 2010, 2016). More recent work uses contextual embeddings to discover economic links for momentum spillover, while related stock-embedding methods combine text with network information (Chung and Tanaka-Ishii, 2023; Takayanagi et al., 2024). Topic-based studies similarly show that peer returns extracted from MD&A narratives or earnings-call business aspects predict focal-firm returns (Zhang et al., 2026; Jin, 2024). These pipelines build on financial text representations, from sentiment lexicons and domain language models (Tetlock, 2007; Loughran and McDonald, 2011; Araci, 2019; Wu et al., 2023) to modern embedding backbones (Reimers and Gurevych, 2019; Xiao et al., 2023). Financial NLP benchmarks cover language understanding and embedding retrieval or similarity (Shah et al., 2022; Jørgensen et al., 2023; Liu et al., 2024; Tang and Yang, 2025). However, these resources score linguistic or document-level quality rather than firm-level return prediction, and most text-derived peer-return studies remain confined to a single market.

D.1 Financial NLP Benchmarks

Financial NLP benchmarks standardise evaluation for sentiment, classification, retrieval, and semantic-similarity tasks on financial text, but none of them ship the firm-level cross-market return-prediction surface CrossAlpha is built around. Araci (2019) introduces FinBERT, a domain-adapted language model for financial sentiment. Shah et al. (2022) releases FLUE, a multi-task benchmark spanning English-language financial text classification. Jørgensen et al. (2023) extends benchmark coverage beyond English financial documents but at the document-classification level rather than firm-level peer linking. Liu et al. (2024) introduce a financial narrative semantic-similarity task for detecting subtle statement-level shifts. Tang and Yang (2025) releases FinMTEB, a massive embedding benchmark covering retrieval and similarity but not return prediction. These resources are valuable model-evaluation suites, but their supervision is primarily linguistic or document-level. They do not ship a firm-level panel that aligns annual reports, daily prices, cross-market peer links, and downstream factor-construction code. A representation that performs well on these NLP tasks may still fail the economic test that matters for asset pricing: selecting firms whose future returns are predictable from economically related firms' disclosures.

D.2 Cross-Market and Graph-Based Return Models

Classic international-finance studies establish that shocks, returns, and volatility are transmitted across major equity markets (Eun and Shim, 1989; Hamao et al., 1990; King and Wadhwani, 1990). Related decompositions show that cross-country stock-market linkages reflect both global financial factors and real economic channels such as trade (Forbes and Chinn, 2004). A more recent line of work moves beyond a single exchange by training prediction models on market time-series features across multiple markets, or by augmenting return prediction with graph structure. Lin et al. (2025) train a cross-market predictor over market

time-series features and show it outperforms single-market baselines. Bukhari et al. (2025) use a graph neural network with intra- and inter-market message passing for cross-market return forecasting. Liu et al. (2026) build a U.S.–China bipartite graph for cross-market return forecasting. Feng et al. (2019) and Kim et al. (2019) are representative domestic relational stock-prediction models: they use company relations or relation-type attention to improve stock ranking and movement prediction. Du et al. (2025) augment a return-prediction architecture with a sparse supply-chain or money-flow graph alongside firm-level features. These methods are complementary to CrossAlpha but do not solve the dataset problem. Price-only cross-market models observe co-movement after the fact but do not explain why two firms should be economically linked, while existing graph datasets either remain domestic or are not released as reusable cross-border annual-report resources. CrossAlpha instead provides text-grounded firm similarity across five markets and a typed, directed US→Japan economic-linkage overlay, enabling both predictive evaluation and relationship-level interpretation.

D.3 Text-Derived Firm Networks and Peer Momentum

Text-derived peer networks are the closest methodological neighbour to CrossAlpha. Hoberg and Phillips (2016) construct text-based network industries from 10-K product descriptions, establishing that disclosure text can recover economically meaningful peer sets. Chung and Tanaka-Ishii (2023) use financial documents to discover economic links for momentum spillover, and Takayanagi et al. (2024) combine textual and network information for stock embeddings. Zhang et al. (2026) construct MD&A topic peers and show that peer lagged returns predict focal-firm returns in the Chinese market. Jin (2024) reaches a similar conclusion from earnings conference-call topic similarity, where business-aspect peers reveal difficult-to-observe firm relatedness. Most directly related, Breitung and Mueller (2025) use LLM-generated historical business descriptions and OpenAI embeddings to construct global business networks and study global-stock lead–lag effects. CrossAlpha builds on this direction but changes the unit of contribution: it releases a multi-market annual-report benchmark with standardized schema fields, dense cross-market firm links, aligned OHLCV data, and reproducible factor protocols, rather than only demonstrating that text-derived business networks are useful in downstream finance applications.

D.4 LLMs and Agents for Financial Extraction

General-purpose LLMs have been used to extract forecast-useful information from financial news and filings. Lopez-Lira and Tang (2023) show that GPT sentiment scores on news headlines predict next-day equity returns in the US. Kim et al. (2024) demonstrate that GPT-4 with chain-of-thought reasoning over earnings releases surpasses human analysts in binary earnings-direction forecasting. On the agentic side, Xiao et al. (2024) use a multi-step LLM workflow that combines news, filings, and price context to issue trading decisions. Fatouros et al. (2025) construct MarketSenseAI, an agent system that explicitly retrieves and reasons over corporate disclosures for market analysis. CrossAlpha uses LLMs differently: not as a trading agent, but as dataset-construction infrastructure. The Distiller standardises heterogeneous annual reports into a shared business schema, and the Linker mines typed, directed economic relationships that are difficult to recover from surface mentions alone. This stores the LLM outputs as reusable data fields that can be evaluated with conventional asset-pricing metrics.

D.5 Supply-Chain Networks in Asset Pricing

Acemoglu et al. (2012) show that idiosyncratic firm-level shocks aggregate into sizeable macroeconomic fluctuations when the production network is sufficiently asymmetric. Herskovic (2018) embed production networks into an asset-pricing model and show that network-implied exposure generates cross-sectional return predictability: firms with similar upstream and downstream connectivity earn correlated excess returns. Empirical identification of supply-chain effects is challenging because supply linkages are typically proprietary; Barrot and Sauvagnat (2016) exploit natural disasters as quasi-experiments, showing that input specificity amplifies shock propagation, with downstream equity prices falling significantly following upstream disruptions. These findings motivate CrossAlpha's Linker pass: by parsing relationship disclosures from US 10-K filings (and cross-reading the Asian counterparties' own filings), we operationalise the production-network channel at scale, achieving a 47× coverage expansion over direct mentions.

E Release, Reproducibility, and Construction Cost

E.1 The `AR-Scraper` Toolkit

To make CrossAlpha a *living* benchmark rather than a one-shot dump, the anonymized release includes `AR-Scraper`, a modular Python toolkit that exposes one scraper module per regulatory system behind a common `fetch(ticker, fy)` interface (Table 29), together with the Distiller prompt used in this work. Researchers can therefore reproduce the full pipeline end-to-end, from raw filings to evaluated factors, on new vintages, new markets, or alternative LLM backends. All data will be released under CC-BY-4.0 and code under MIT.

Table 29: The five `AR-Scraper` modules shipped with CrossAlpha. Module names are suffixes under the `ar_scraper` package; each implements the same `fetch(ticker, fy)` interface.

Module	Market	System	Lang.	Doc type
`edgar`	US	SEC EDGAR	en	10-K
`edinet`	JP	EDINET	ja	*yuho*
`mops`	TW	MOPS	zh-Hant	annual report
`dart`	KR	DART	ko	business report
`hkex`	HK	HKEx	en / zh-Hant	annual report

E.2 Construction Cost and Latency

LLM spend per dataset refresh. All figures are computed from public pricing tiers as of early 2026. (1) *Distiller pass* uses GPT-4.1 at \$2 / \$8 per million input / output tokens, averaging ≈40k input and ≈2k output tokens per filing, for ≈\$0.10 per filing or ≈**\$1,000 total** across ≈10,700 filings. (2) *Dense embedding* uses OpenAI `text-embedding-3-large` at \$0.13/M tokens; ten category strings per filing at ≈2k tokens each cost ≈**\$5 total**. (3) *Linker pass* (US→JP typed-edge layer) uses GPT-4.1 on ≈9,200 candidate pairs at ≈5k input and ≈0.5k output tokens per call, for ≈**\$25 total**; the open-universe retriever variant adds another ≈**\$30**. (4) The *event-tracing* LLM filter (Section 5.3) uses a single GPT-5 call per source stock on ≈60 neighbours (graph+typing book) or on the full ∼2,000-firm universe (agent-solo ablation) for 25 source stocks at a fixed cost of ≈**\$8 total**; the typed output is the per-source-stock annotation we release as part of the event-tracing benchmark. Together, a full CrossAlpha refresh costs about **\$1,070** in API spend (Distiller \$1,000 + embedding \$5 + Linker \$25 + retriever \$30 + LLM filter \$8, rounded), a one-time inference budget for the entire cross-market benchmark.

Wall-clock latency. Respecting provider rate limits with ≤ 20 concurrent requests, the Distiller pass completes in ≈8 hours and embedding in <1 hour. Backtesting the monthly factor harness is CPU-bound and completes in ≈30 minutes on a single workstation. The base dense benchmark can therefore be refreshed within a working day.

API-cost scope for users. Table 30 separates the cost of using the benchmark from the cost of rebuilding it. Using the frozen benchmark and reproducing every reported evaluation costs nothing in API spend; distilling all ≈10,700 filings with the hosted open model of Appendix C.14 is estimated at ≈\$71 by linear extrapolation from the 600-filing diagnostic (core distillation only, excluding the optional Linker, retriever, and event-filter refreshes).

Table 30: API-cost scope.

Usage	Cost
Use the frozen benchmark and reproduce its evaluations	\$0
Qwen Distiller diagnostic, 600 filings	≈\$4
Linear estimate for distilling all 10,700 filings with the hosted open model	≈\$71
Original full benchmark refresh with the documented proprietary models	≈\$1,070

F Distiller and Linker Prompts

This appendix gives the prompt templates used to instantiate the two role-specialised LLM passes introduced in Section 3.1: the **Distiller** (ϕ_D, GPT-4.1) for market-level filing standardisation and the **Linker** (ϕ_L, GPT-4.1, same backbone, relational-reasoning role) for typed-edge mining. Both prompts use temperature 0 and require strict JSON output; field names match those used in tables and figures throughout the main text. The full reference implementation, including parsing helpers, retry logic, and per-market wrappers, ships with the `AR-Scraper` toolkit (Appendix E.1).

F.1 Distiller Prompt (ϕ_D): 10-Category Standardisation

Each filing is processed independently. The system message is shared across all five markets; only the user message changes per filing. Output tokens are capped at 2,000 to control cost; on truncation we retry the offending field individually.

Distiller system message (ϕ_D, GPT-4.1)

You are an expert multilingual financial analyst fluent in English, Japanese, Chinese (Traditional and Simplified), and Korean. Your task is to read a corporate annual report (10-K, yuho, business report, or equivalent) and distil it into ten structured English business-description fields.

Rules:
1. Output English only, even when the source filing is in another language.
2. Be concise: each field $\leq$ 80 tokens.
3. Use only facts present in or directly entailed by the filing; do not import outside knowledge.
4. Avoid boilerplate (risk-factor language, forward-looking caveats, generic mission statements).
5. If a field cannot be inferred from the filing, output `null`.
6. Respond ONLY with a single JSON object; no explanatory text, no markdown.

Distiller user message template (ϕ_D)

TICKER: {{ticker}}
MARKET: {{market}}
FISCAL_YEAR: {{fy}}
FILING:
{{filing_text}}

Return a JSON object with EXACTLY these ten keys (string values, English):
1. `main_business_segments`
2. `core_technologies`
3. `primary_customers`
4. `supply_chain_position`
5. `geographic_coverage`
6. `financial_profile`
7. `revenue_model`
8. `value_proposition`
9. `strategic_focus`
10. `key_competitors`

F.2 Linker Prompt (ϕ_L): Typed Cross-Market Edge Mining

For the US→JP source–target setting, the Linker is invoked once per US 10-K with the candidate JP universe (TOPIX 500 tickers + GICS sectors) injected into the user message. The dual-side reading mechanism is encoded by allowing the agent to mark an edge as supported by `supplier_side_disclosure`, i.e., the edge is recovered from the *target's yuho* rather than the source 10-K. For instance, Apple's 10-K names no Japanese firm (only generic "outsourcing partners in Asia"), yet reading the targets' *yuho* yields typed supplier edges from Apple to Sony, Murata, TDK, and Advantest; these links are not explicit in Apple's own filing.

Linker system message (ϕ_L, GPT-4.1)

You are an expert in global supply chains, corporate disclosure, and cross-border equity research. You will be given (a) a US 10-K filing and (b) a target foreign market. Your task is to enumerate every named or strongly inferable foreign firm in the target market that has one of four typed relationships to the source firm:

`SUPPLIER` : provides inputs, components, or services to the source.
`CUSTOMER` : purchases material volume from the source.
`PARTNER` : joint venture, alliance, or co-development.
`COMPETITOR`: directly overlapping product or service market.

Rules:
1. Combine explicit text mentions in the source 10-K with grounded world knowledge (e.g., a documented major customer relationship reported in the supplier's own filing).
2. When the 10-K is silent but the relationship is well-attested in the target firm's public disclosures, mark `evidence = "supplier_side_disclosure"` and lower the confidence accordingly.
3. Do NOT include US-domestic firms; relations must cross the source/target border.
4. Provide a one-sentence `rationale` naming the specific product, segment, or contract that grounds the edge.
5. Respond ONLY with strict JSON.

Linker user message template (ϕ_L)

SOURCE_TICKER: {{us_ticker}}
SOURCE_FILING (10-K):
{{us_10k_text}}

TARGET_MARKET: {{market}}
TARGET_UNIVERSE (ticker, GICS sector):
{{candidate_table}}

Return a JSON list. Each element:
{
"target_ticker": string,
"relation_type": "supplier" | "customer" | "partner" | "competitor",
"evidence": "explicit_mention" | "inferred_world_knowledge" | "supplier_side_disclosure",
"confidence": float in [0, 1],
"rationale": string ($\leq$ 40 words)
}

Static-edge-mining variants. The Linker variants that produced the typed-edge release reuse the system message above with the following changes to the user message: ϕ_L^{A} (per-pair grader) is invoked once per (US, JP) candidate pair from the embedding top-20 and outputs a single edge with a continuous strength score $s \in [0, 1]$; ϕ_L^{B} (comparative reranker) receives all $K{=}50$ embedding candidates in one prompt and returns a ranked top-10 with per-pair strengths; ϕ_L^{C} (retriever) drops the candidate table and instead exposes a `search(query)` tool over the full Russell 1000.

F.3 Event-Tracing Prompt (ϕ_L^{event}): Typed Sign on a Hub Event

The event-tracing prompt of Section 5.3 is invoked *once per source stock* (not per pair), takes as input the list of the source stock's top-K dense-graph cross-market neighbours with their identity fields, and returns a typed sign + confidence + relation for each. The system message instantiates the source-stock-specific knowledge in-place (here SK Hynix; one would swap in the source stock's identity for a different test).

Event-tracing system message (ϕ_L^{event}, GPT-4.1, temperature 0)

You are a sell-side equity analyst who reasons about cross-market spillover. SK Hynix (000660.KS, South Korea, semiconductor / DRAM & HBM memory manufacturing) has just had a LARGE POSITIVE IDIOSYNCRATIC stock move (a memory-demand or memory-pricing surprise, the typical SK Hynix shock in the 2023-26 AI cycle). For each firm in the list, predict the sign of its ABNORMAL (market-relative) stock return over the NEXT 1-3 trading days, reasoning about the economic link:

sign = +1 – co-mover: semiconductor-equipment / materials supplier to memory makers; memory or broad-semis sector peer; AI-accelerator or HBM customer whose demand is validated; fab / foundry whose capex outlook improves.

sign = -1 – inverse mover: a firm whose MAJOR INPUT COST is DRAM/NAND/HBM memory and has little offsetting upside (e.g. a pure PC / server / handset / consumer-electronics OEM that BUYS memory).

sign = 0 – no material economic exposure to a SK Hynix memory shock.

Respond ONLY with a JSON list, one object per firm, in the SAME ORDER as the input:
{"ticker": "...",
"relation": "<supplier|customer|sector_peer|foundry|memory_consumer|unrelated|...>",
"sign": -1|0|1, "confidence": <float 0-1>,
"rationale": "«=25 words>"}

Event-tracing user message template (ϕ_L^{event})

FIRMS (SK Hynix's CrossAlpha embedding neighbours, in similarity-rank order):
- {{ticker}} | {{name}} | {{country}} | industry: {{industry}} | GICS: {{gics}} | (market: {{region}})
- ... (one line per neighbour, global top-$K = 60$ pooled across the three non-home markets)

The benchmark protocol scores any future LLM filter by running its labelled-sign output through the same long-only-confident-+ backtest of Table 21 on the same 25-source-stock / 606-event GPT-5 post-cutoff window, so improvements are measured in downstream predictive performance, not in label-matching against a noisy gold set.

G Grounding the 10-Category Schema Across Regulatory Systems

The ten Distiller categories (main_business_segments, core_technologies, primary_customers, supply_chain_position, geographic_coverage, financial_profile, revenue_model, value_proposition, strategic_focus, key_competitors) are chosen to satisfy two constraints. First, each category corresponds to a cross-firm channel that prior firm-level return-predictability research finds informative (Table 31). Second, every category is recoverable from *every* mandatory annual filing across the five regulatory systems we cover (Table 32). This cross-system grounding is what makes the Distiller's schema a true lowest-common-denominator: no category depends on disclosure conventions unique to a single market.

Practical use. Table 32 records where each schema field is obtained across markets. Where a regulator uses sub-sections rather than a single dedicated section (e.g., *primary customers* appears in both Part II.1 and the risk-factor section of a Japanese *yuho*), the Distiller prompt (Appendix F.1) asks the model to merge the relevant fragments into a single concise English description. In the 200-filing audit, 89% of human-audited fields are rated "accurate" (Appendix C.10), with the lowest source-language group at 82%.

Table 31: **The ten Distiller categories and why each is included.** Each category captures a cross-firm channel that firm-level return-predictability research finds informative; the full set is the intersection of these channels with content mandated across all five filing systems.

Category	Cross-firm channel it captures	Motivated by
`main_business_segments`	Product-market and industry membership	Hoberg and Phillips (2016)
`revenue_model`	Business-model similarity	Hoberg and Phillips (2010); Bini et al. (2016)
`value_proposition`	Product differentiation	Hoberg and Phillips (2010)
`primary_customers`	Customer lead–lag links	Cohen and Frazzini (2008)
`supply_chain_position`	Production-network position	Acemoglu et al. (2012); Herskovic (2018); Barrot and Sauvagnat (2016)
`key_competitors`	Direct competitor links	Hoberg and Phillips (2016)
`core_technologies`	Shared technology and R&D exposure	Bloom et al. (2013); Bybee et al. (2024)
`strategic_focus`	Forward strategy and R&D direction	Muslu et al. (2015); Bybee et al. (2024)
`geographic_coverage`	Common geographic and market exposure	García and Norli (2012); Menzly and Ozbas (2010)
`financial_profile`	Size and growth characteristics	Fama and French (2015); Harvey et al. (2016)

Table 32: Each Distiller category maps to a mandatory disclosure surface in all five regulatory systems. References use each regulator's canonical section identifier; bracketed phrases give the local-language section title for the non-English systems.

Distiller category	SEC 10-K (US)	JP *yuho* (Japan)	KR DART (Korea)	TW MOPS (Taiwan)	HKEx AR (Hong Kong)
`main_business_segments`	Item 1	yuho Part II.1 [*Business*]	II.1 [*Business overview*]	§3 [*Business*]	MD&A: Business
`core_technologies`	Item 1	yuho Part II.2 [*R&D*]	II.5 [*R&D*]	§5 [*R&D*]	MD&A: R&D
`primary_customers`	Item 1	yuho Part II.1 [*Customers*]	II.1 [*Major customers*]	§3 [*Customers*]	MD&A: Customers
`supply_chain_position`	Item 1	yuho Part II.1 [*Suppliers*]	II.4 [*Raw materials*]	§4 [*Raw materials*]	MD&A: Supply
`geographic_coverage`	Item 1	yuho Part I.3 [*Establishments*]	I.4 [*Sites*]	§3 [*Operations*]	MD&A: Geog.
`financial_profile`	Item 6/8	yuho Part V [*Financials*]	III [*Financials*]	§6 [*Financials*]	Fin. Stmts
`revenue_model`	Item 7	yuho Part II.3 [*Performance*]	II.2 [*Sales*]	§3 [*Revenue*]	MD&A: Rev.
`value_proposition`	Item 1	yuho Part II.1 [*Strengths*]	II.1 [*Comp. advantage*]	§3 [*Comp. advantage*]	MD&A: Strategy
`strategic_focus`	Item 7	yuho Part II.2 [*Challenges*]	II.6 [*Strategy*]	§5 [*Strategy*]	MD&A: Outlook
`key_competitors`	Item 1	yuho Part II.1 [*Competition*]	II.1 [*Competitors*]	§3 [*Competitors*]	MD&A: Comp.