Does Language Matter in LLM-Based Business Networks? Evidence from Korean Corporate Disclosures*

Seon Ah Lee** (Sungkyunkwan University)
Byung Hwa Lim*** (Sungkyunkwan University)

〈Abstract〉

This study constructs large language model (LLM)-based dynamic business networks for KOSPI-listed firms and examines whether the output language of LLM-generated business descriptions affects network informativeness. Using annual reports from 2010 to 2024, we extract the Business Overview and MD&A sections and generate standardized business descriptions in both Korean and English from the same underlying disclosures. To reduce look-ahead bias, firm and product identifiers are masked before LLM generation. We then embed the generated descriptions with multiple language and finance-specific models and construct annual peer networks based on cosine similarity. The results show that the proposed networks identify firm-level economic peers not fully captured by the Korean Standard Industrial Classification (KSIC). More importantly, Korean summaries and English summaries generated from Korean disclosures produce different similarity structures and lead-lag return signals. Asset-pricing tests indicate that network-based peer-return factors generate significant alphas beyond standard risk factors and a KSIC-based benchmark. LLM-based English business descriptions tend to deliver more stable and broadly significant lead-lag performance in several specifications, while Korean descriptions also provide meaningful signals in selected cases. These findings suggest that output-language choice is not a neutral preprocessing step, but a substantive modeling decision in LLM-based financial text analysis.

* This article is a revised and condensed version of the first author's master's thesis. This research was supported by the "Regional Innovation System & Education (RISE)" through the Seoul RISE Center, funded by the Ministry of Education (MOE) and the Seoul Metropolitan Government (2026-RISE-01-018-04).
JEL Classification: G12, G14, C55, L16
** First author, Department of Fintech, Sungkyunkwan University, E-mail: dltjsdk0321@gmail.com
*** Corresponding author, Department of Fintech, SKK Business School, Sungkyunkwan University, E-mail: limbh@skku.edu

Keywords: Industry Classification, Business Network, Large Language Models, Textual Analysis, Lead-lag Effect
Article History: Received 9 May 2026, Revised 3 July 2026, Accepted 12 July 2026

[1] Introduction

Industry classification plays a central role in financial research and practice. It provides a framework for identifying economically comparable firms, measuring common risk exposures, constructing peer groups, and analyzing information diffusion across firms. In asset pricing and portfolio management, industry classifications are also widely used to define benchmark portfolios, control for sectoral risk, and interpret cross-sectional return patterns. For these purposes, an industry classification system should capture the economic proximity among firms whose business activities, product markets, and sources of cash flows are substantively related.

However, conventional industry classification systems face increasing limitations as firms expand across multiple business areas and as technological convergence blurs traditional industry boundaries. The Korean Standard Industrial Classification (KSIC), like many standard classification systems, assigns each firm to a single representative industry category. Although such a single-label classification structure provides stability and administrative clarity, it may fail to reflect the economic substance of diversified firms. Broad industry codes can also generate classification-induced transitivity: firms grouped under the same category may be treated as comparable even when their actual business activities differ substantially.

These limitations have motivated research that measures business relatedness directly from textual firm descriptions. Hoberg and Phillips (2010, 2016) show that text-based measures of product-market similarity derived from firms' business descriptions can capture economically meaningful peer relationships beyond static industry classifications. More recently, advances in large language models (LLMs) have made it possible to standardize heterogeneous corporate disclosures into structured business descriptions before measuring similarity. Breitung and Müller (2025) provide an important recent example by showing that LLM-generated business descriptions can be used to construct economically meaningful business networks in a broad international setting. However, their setting does not directly examine whether, when the underlying disclosures originate in a non-English language, the output language of LLM-generated business descriptions changes the resulting similarity structure. In this respect, firm-level evidence from Korean corporate disclosures remains

limited. Existing Korean-market studies suggest that textual information can contain economically relevant pricing signals and that its empirical usefulness may vary with the construction of text-based measures. Woo and Kim (2017) document a significant relation between online text and stock-price movements in Korea, while Lee and Park (2021) show that predictive performance differs across text-derived signals. These findings support the use of Korean corporate disclosures as a firm-level textual source for measuring economic relatedness.

This study extends that literature to the Korean market and focuses on a question that is particularly relevant in a non-English disclosure environment. The original corporate disclosures in our setting are written in Korean, whereas many frontier LLM and embedding systems have been developed primarily in English or in multilingual settings with uneven language coverage. This raises the possibility that, even when the same source disclosure is used, Korean summaries and English summaries generated from Korean disclosures may produce different similarity structures, peer assignments, and return-predictive signals.

The central research question of this paper is whether the output language of LLM-generated business descriptions affects the economic informativeness of the resulting business networks. This question matters because, if output-language choice changes similarity measurement and peer identification even when the underlying disclosures are fixed, it can also affect downstream applications such as industry classification, peer-group construction, and return predictability. To address this question, we construct LLM-based dynamic business networks for KOSPI-listed firms using the "Business Overview" and "Management's Discussion and Analysis" sections of annual business reports filed through DART from 2010 to 2024. Extending the text-based network approach of Hoberg and Phillips (2010, 2016), we generate both Korean summaries and English summaries from the same underlying Korean disclosures using the same source documents, substantively equivalent prompt instructions, and the same generation settings, varying only the output language of the LLM-generated business descriptions. We then embed the generated descriptions, measure pairwise business similarity using cosine similarity, and construct annual peer networks that are not restricted by pre-assigned industry boundaries. To mitigate look-ahead bias, firm and product identifiers are masked before LLM generation.

We evaluate the proposed networks in three ways. First, we examine whether they capture stable and economically interpretable business proximity over time. Second, we assess whether they provide a more informative firm-level representation of business relatedness than KSIC. Third, we evaluate the networks from an asset-pricing perspective through peer-based lead-lag tests. This design is motivated by the literature on information diffusion in stock returns. Lo and MacKinlay (1990) document lead-lag effects from large to small firms, while Hou (2007) shows that such effects are concentrated among economically related firms. Related evidence is also found in the Korean market (Park and Jang, 2003; Kang and Yoon, 2012; Park and Heo, 2020). Taken together, this literature suggests that the quality of peer

identification is important for understanding return spillovers across firms. Building on this insight, we test whether peer-return signals based on the proposed business network generate significant alphas after controlling for standard risk factors and a KSIC-based benchmark.

The empirical results show that the proposed networks identify economically meaningful peer relationships not fully captured by KSIC and generate significant lead-lag alphas in several specifications. More importantly, Korean summaries and English summaries generated from the same Korean disclosures produce different similarity structures and return-predictive signals. English summaries generated from Korean disclosures tend to deliver more stable and broadly significant lead-lag performance in several settings, while Korean summaries also provide meaningful signals in selected cases.

This study contributes to the literature in three ways. First, it provides firm-level evidence on LLM-based business-network construction using Korean corporate disclosures. Second, it directly compares Korean summaries and English summaries generated from the same Korean source disclosures, thereby identifying the effect of output-language choice under otherwise comparable conditions. Third, it shows that these language-dependent differences are economically relevant by linking them to business-similarity measurement, peer identification, and asset-pricing performance.

The remainder of the paper is organized as follows. Section 2 describes the data and factor construction. Section 3 explains the LLM-based business description procedure, look-ahead bias mitigation, and business-network construction methodology. Section 4 evaluates the proposed networks in terms of temporal stability and asset-pricing performance. Section 5 concludes.

[2] Data

2.1 Business description

This study collected annual business reports filed through the DART (Data Analysis, Retrieval and Transfer System) for KOSPI-listed firms from 2010 to 2024. We focus on KOSPI-listed firms because they account for a substantial share of the Korean equity market and are subject to relatively standardized disclosure practices. To reduce survivorship bias, the sample includes all firms that were listed in each year. As a result, the number of firms analyzed per year ranges from 614 to 793, with an annual average of approximately 671 firms. We obtained the reports through the DART API, using each firm's corp code as a unique identifier. Using an HTML parser, we extracted the textual contents of the reports and retained two sec-

tions for analysis: "Business Overview" and "Management's Discussion and Analysis (MD&A)." The Business Overview section provides detailed information on a firm's core products and services, while the MD&A section provides management's assessment of the firm's operations and strategy.

Because annual business reports are publicly disclosed with a reporting lag, reports filed between July of Year t and June of Year t+1 are treated as business description data for Year t. For benchmarking purposes, we use the Korean Standard Industrial Classification (KSIC) provided by the Korea Exchange (KRX). To maintain temporal alignment with the business report data, we obtain point-in-time KSIC classifications as of the end of June in Year t+1. These KSIC classifications, together with the factors constructed from the business description data, are then used for portfolio rebalancing and comparative analysis over the period from July of Year t+1 to June of Year t+2.

To ensure accurate data linkage across sources, we apply a multi-step verification procedure. First, the annual business reports are matched with the KSIC data using each firm's stock ticker for the corresponding year. To address potential identification discrepancies arising over time, the matched observations are further cross-validated using firm names during the longitudinal matching process.

2.2 Factor Construction

We evaluate the proposed business network using peer-based lead-lag effects and compare its predictive performance with established asset-pricing factors. As a benchmark, we employ a seven-factor model consisting of the five factors in Fama and French (2015), augmented with momentum and short-term reversal factors following Jegadeesh and Titman (1993) and Lo and MacKinlay (1990), respectively. Rather than relying on the standard FnGuide factor series, which are based on the combined KOSPI and KOSDAQ universe, we construct the benchmark factors separately for the KOSPI universe.

The data are obtained from FnGuide. Specifically, we use monthly total returns adjusted for dividends, market capitalization, accounting variables required for factor construction, market returns, and the 91-day certificate of deposit (CD) rate as the risk-free rate. The accounting variables include book equity, total assets, revenue, cost of goods sold, selling, general and administrative expenses, and interest expenses. The financial dataset is linked to the annual report sample using both firm names and stock tickers.

Mkt is defined as the excess return on the market portfolio over the risk-free rate. SMB and HML are constructed from 2×3 sorts on size and book-to-market equity, while RMW and CMA are based on 2×3 sorts on size with operating profitability and investment, respectively. MOM and REV are constructed from 2×3 sorts on size with prior 2-to-12-month returns and prior one-month returns, respectively. Following the standard convention, accounting-based factors are rebalanced annually from July of Year t to June of Year t+1, whereas return-based factors are updated monthly.

3 Methodology

3.1 LLM-based business descriptions

The annual business reports collected in Section 2.1 vary substantially in length across firms and contain a mixture of narrative disclosures and tabular information. Such heterogeneity can induce spurious similarity in embedding-based cosine similarity measures, as models may respond to formal document characteristics, such as document length, reporting format, writing style, repeated expressions, and the presence of extensive tables, rather than underlying business content. To improve semantic comparability across firms, we generate standardized business descriptions from the annual reports using a large language model (LLM). Specifically, we employ the Google/Gemini 2.5 Flash Lite API for text generation, given its suitability for large-scale batch processing and its cost-efficient pricing structure.

The prompt design was guided primarily by the need to standardize output length across firms, while also taking into account the input-length constraints of the embedding models used in the subsequent analysis. Generated descriptions were constrained to 400-512 tokens, which helps reduce firm-size-related variation in disclosure length and improves comparability across firms. This design choice is particularly important because the original annual reports exhibit substantial variation in length across firms of different sizes. The prompt also instructed the model, whenever possible, to identify the business segment accounting for the largest share of revenue so that the firm's core business model would be clearly reflected. In addition, the model was instructed to rely exclusively on the information contained in the given annual report in order to minimize the risk of look-ahead bias.

The same prompt template and generation settings were applied to all firms and years. Generation was conducted at a temperature of 0.0 to encourage consistent and objective outputs. If a generated description fell outside the target token range, the model was prompted again with an additional instruction indicating whether the previous output was shorter than 400 tokens or longer than 512 tokens, together with the realized token count. Because the annual reports include both narrative text and numerical disclosures presented in tables, the prompt was designed to integrate qualitative business descriptions with quantitative information extracted in text form from tabular contents.

Both Korean and English summaries were generated from the same underlying Korean disclosures using Korean- and English- language prompt templates that contained substantively equivalent instructions and the same generation settings. The full prompt templates used for the English and Korean summaries are provided in Appendix A. The resulting summaries were then evaluated across different embedding models. Table 1 presents an example of the LLM-generated business descriptions for 2024 in Korean and in English, showing that

the two outputs are broadly aligned in their depiction of the firm's core business while exhibiting minor differences in emphasis and expression.

〈Table 1〉 LLM-based Business Description Example

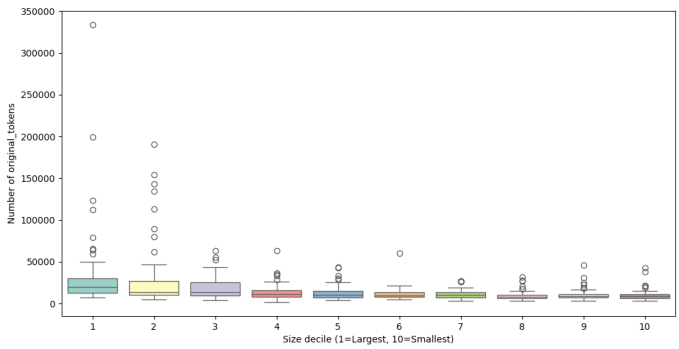
This table presents the LLM-generated English and Korean business descriptions for AJ Networks in 2024. AJ Networks is classified under the KSIC large category of 'General Services'. Both descriptions identify rental as the firm's core business while also reflecting related activities in warehouse leasing, distribution, IT-related services, and real-estate management. This example shows that the generated descriptions provide a more detailed representation of firm-level business activities than the broad KSIC category alone.

AJ Networks	
English Version	Korean Version
The company operates a diversified business model structured into three primary segments: Rental, Warehouse & Distribution, and Other. The Rental segment is the largest contributor to total revenue, consistently accounting for over 60% of consolidated sales in the fiscal years 2022, 2023, and 2024. This segment encompasses the rental and distribution of various assets, including logistics equipment such as pallets, industrial safety equipment like aerial work platforms and forklifts, and IT devices such as OA equipment, mobile devices (tablets, smartphones), environmental equipment (air purifiers), and robots (serving and collaborative robots). The company targets a broad range of corporate clients, financial institutions, and public organizations, and holds a dominant market share in the aerial work platform rental sector. The Warehouse & Distribution segment represents the second-largest revenue stream, contributing approximately 34% of total consolidated revenue in fiscal year 2024. This segment includes cold storage and freezer warehouse leasing operations in Vietnam, as well as fuel retail and solar power generation through its subsidiary. Additionally, it involves the distribution of hardware, including servers and storage devices, alongside IT outsourcing services and contact center operations. The Other segment, comprising real estate management services, accounts for the smallest portion of revenue, around 3.5% in fiscal year 2024. In fiscal year 2024, the company reported consolidated revenue of KRW 1,014,156 million, an 8.29% increase year-over-year, driven by the bal-	본 기업의 핵심 비즈니스 모델은 다양한 자산을 렌탈하고 유통하는 사업을 중심으로 구성되어 있습니다. 사업보고서에 따르면, 2024년 연결 기준 매출액 1조 142억원 중 렌탈 부문이 6,337억 3,600만원으로 62.5%를 차지하며 가장 높은 비중을 보였습니다. 이는 파렛트 등 물류기자재, OA기기, 모바일기기, 환경기기, 로봇, 고소장비, 지게차 등 광범위한 품목의 B2B 렌탈 및 유통 사업을 통해 수익을 창출하고 있음을 시사합니다. 창고 및 유통부문은 3,447억 7,000만원(34.0%)의 매출을 기록하며 두 번째로 큰 비중을 차지했습니다. 이 부문에서는 베트남 냉장 ㅇ 냉동 창고 임대업, 유류 도소매 및 태양광 발전 사업, 서버 등 저장장치 유통 및 IT 아웃소싱 서비스 등을 영위하고 있습니다. 기타부문은 부동산 관리 서비스 등을 포함하며 356억 5,000만원(3.5%)의 매출을 기록했습니다. 이사의 경영진단 및 분석의견에서도 렌탈사업의 중요성이 강조되며, 렌탈자산 다변화를 통한 시장 지배력 유지와 로봇 렌탈 등 신규 사업 확장에 대한 의지가 나타나 있습니다. 또한, 2024년 매출액은 전년 대비 8.29% 증가했으나, 영업이익은 판관비 증가 및 시스템 도입 등으로 7.24% 하락한 것으로 분석되었습니다. 전반적으로 기업은 핵심 렌탈 사업의 경쟁력을 유지하면서 미래 성장 동력 확보를 위한 노력을 병행하고 있습니다.

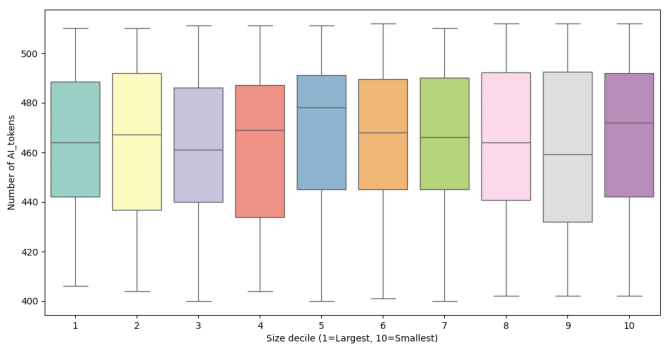
anced growth across its business units. However, operating profit decreased by 7.24% to KRW 72,790 million, attributed to increased selling, general, and administrative expenses and the implementation of a new system. The company is actively pursuing strategies to enhance its competitive edge, including expanding its rental asset portfolio, focusing on new business areas like robot rentals, and exploring mergers and acquisitions.	

Figure 1 compares the token counts of the original annual reports, defined as the combined length of the 'Business Overview' and 'MD&A' sections, with those of the LLM-generated English business descriptions for 2024. Whereas the original reports exhibit substantial dispersion in token counts across firms, the generated descriptions maintain a much more consistent length, with an average of approximately 457 tokens. This indicates that the LLM-based standardization substantially reduces cross-firm variation in document length.

〈Figure 1〉 Comparison of Business Description Token Length



(a) Original Business Description from DART Annual Reports



(b) LLM-based Business Description

This figure compares token counts across firm-size deciles for the original DART business descriptions and the LLM-generated business descriptions in 2024, showing that the latter are much more concentrated around the target range.

3.2 Mitigating Look-ahead Bias

Generating business descriptions with LLMs raises the possibility of look-ahead bias if the model draws on pre-trained knowledge that is not contained in the source report. Glasserman and Lin (2023) show that company identifiers may activate model knowledge unrelated to the contemporaneous text, implying that anonymization can help mitigate this source of bias. To reduce this risk, we masked firm names and product names in the original annual-report texts before generating the LLM-based business descriptions.

For the initial masking step, we used the spaCy named entity recognition model (ko_core_news_lg) to detect firm and product names and replace them with standardized placeholders such as "Company ABC" and "Product ABC." However, removing firm and product names alone does not necessarily eliminate all identifying information, because firms may still be inferable from indirect business-specific cues such as distinctive technologies, market-share descriptions, flagship business segments, or other highly specific operational details. To address this issue, we implemented an additional verification step. Specifically, we provided the masked texts to gpt-4o-mini and asked the model to infer the firm's identity if possible. When the model correctly identified the firm name, the corresponding report was reviewed manually and additional masking was applied to words or expressions that could plausibly reveal the firm's identity. The GPT-based verification step initially flagged approximately 20–30% of firm-year observations as potentially identifiable. These flagged observations were then manually reviewed, and additional masking was applied before generating the final business descriptions. Thus, the 20–30% figure refers to the initial flagging rate during the verification process, not to the residual identifiability of the final masked inputs.

3.3 Constructing business networks

Business similarity is measured using cosine similarity computed from the textual descriptions of firms' business activities, following Hoberg and Phillips (2010, 2016). The embedding models were selected to span three dimensions that are central to our empirical setting. First, we compare general-purpose representations with finance-specialized representations by including BERT/KoBERT and FinBERT/KR-FinBERT. This choice is motivated by prior evidence that domain-specific pretraining can improve performance on financial text (Yang et al., 2020). Second, because the source disclosures are written in Korean while the LLM-generated business descriptions are analyzed in both Korean and English, we include model families that differ in Korean-language adaptation and broader multilingual or English-centered representation capacity. Third, we compare traditional transformer-based encoder embeddings with recent large-scale embedding models by including Google/Gemini-embedding-001 (Gemini) and OpenAI/text-embedding-3-large (GPT-L). These models are also closely related to the recent literature on LLM-based business descriptions and global business-network con-

struction and have shown strong performance in recent embedding applications (Breitung and Müller, 2025; Lee et al., 2025). This design allows us to assess whether the informativeness of the resulting business network varies with domain specialization, language compatibility, and embedding architecture.

For the transformer-based encoder models, business descriptions were tokenized with truncation at 512 tokens and converted into document vectors using attention-mask-based pooling. For BERT and FinBERT, document vectors were constructed by mean-pooling the last hidden-state representations, and L2 normalization was applied before similarity calculation. For the API-based embedding models, Gemini embeddings were obtained from gemini-embedding-001 with the task type set to semantic similarity, and GPT-L embeddings were obtained from text-embedding-3-large. Because the generated business descriptions were standardized to a relatively short range of 400–512 tokens, we used a common embedding dimension of 768 for the API-based specifications to maintain comparability across models while preserving sufficient semantic information for similarity measurement. Pairwise business similarity was then computed using cosine similarity between firm-level embedding vectors, so that firms with more closely aligned semantic representations received higher similarity scores.

We use each model to embed the LLM-generated business descriptions and compute annual pairwise cosine similarities across firms. Peer firms are then identified using annual cross-sectional percentile cutoffs over the full distribution of firm-pair similarities. Specifically, we repeat the analysis using the top 1%, 2%, 5%, and 10% of firm pairs as peer links. Because peers are defined using annual percentile cutoffs over all firm pairs, some firms may not be assigned any peers in a given year.

Table 2 summarizes the cosine similarity distributions and network characteristics. Panel A reports cosine-similarity statistics by output language and embedding model. The distributions differ meaningfully across models. GPT-L produces the lowest mean similarity and the widest dispersion for both the English summaries generated from Korean disclosures and the Korean summaries, indicating a more selective similarity structure. By contrast, BERT- and FinBERT-based embeddings assign uniformly higher similarity scores with narrower dispersion, implying a more compressed similarity distribution. Gemini lies between these two extremes. These statistics should be interpreted as descriptive evidence on how alternative embeddings structure firm-level business relatedness, rather than as direct measures of model superiority.

Panel B reports the network characteristics implied by the top 5% peer-selection rule. The average number of peer firms per firm is broadly similar across specifications, but the median is consistently lower than the mean, indicating that peer links are distributed unevenly across firms. The share of firms with no peers also varies across model-language combinations. Gemini yields the lowest proportion of such firms in both languages, whereas BERT produces the highest proportion, especially for Korean summaries. These results show that even under the same top 5% threshold, the structure of the resulting business network var-

ies meaningfully with the output language of LLM-generated business descriptions and the embedding model.

These differences matter because they affect the composition of peer groups used in the temporal-stability, cross-sectional, and asset-pricing analyses that follow.

〈Table 2〉 Descriptive Statistics of LLM-based Business Networks

This table reports descriptive statistics for each LLM-based business network by output language and embedding model, with all statistics averaged over the full sample period. Panel A presents cosine-similarity statistics(mean cosine similarity, median cosine similarity, standard deviation of cosine similarity) and Panel B reports the corresponding network characteristics under the top 5% peer-selection rule.

Panel A. Cosine Similarity Statistics

Language	Embedding Model	Mean	Median	SD
English	GPT-L	0.595	0.593	0.072
	Gemini	0.840	0.839	0.024
	FinBERT	0.915	0.916	0.025
	BERT	0.938	0.941	0.019
Korean	GPT-L	0.541	0.538	0.071
	Gemini	0.830	0.830	0.026
	FinBERT	0.969	0.970	0.008
	BERT	0.938	0.941	0.021

Panel B. Network Characteristics

Language	Embedding Model	Mean Number of Peers	Median Number of Peers	Number of No-Peer Firms (Monopoly)	Share of No-Peer Firms (%)
English	GPT-L	34.19	27.33	14.4	2.12
	Gemini	33.80	26.50	6.8	1.01
	FinBERT	34.03	23.20	11.3	1.69
	BERT	34.70	22.53	24.1	3.62
Korean	GPT-L	34.14	26.73	13.3	1.99
	Gemini	33.84	25.93	7.7	1.15
	FinBERT	34.62	23.20	22.5	3.40
	BERT	35.93	23.23	46.0	6.89

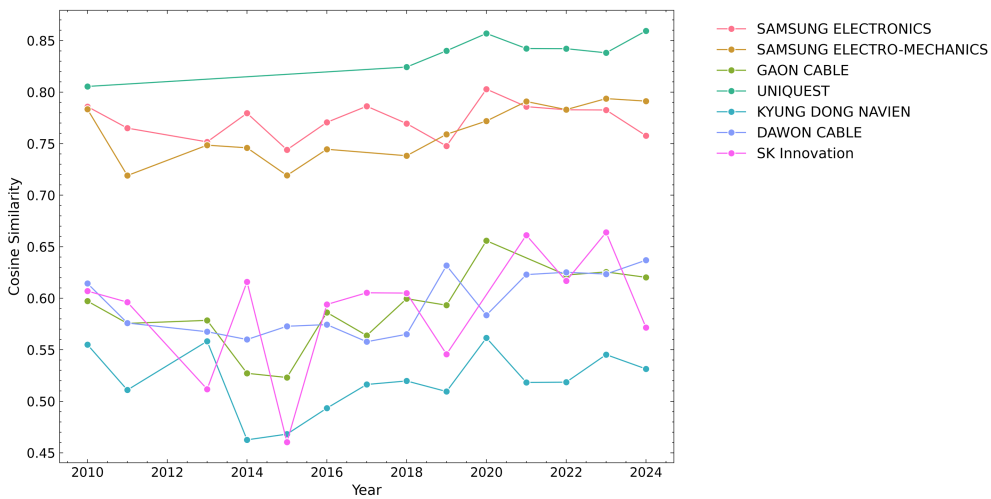
[4] Evaluation of the Business Networks

4.1 Temporal stability of business net works

To illustrate the temporal stability of the proposed business network, we examine the time-series behavior of cosine similarity between two focal firms and a selected set of comparison firms, and contrast these patterns with peer relationships implied by KSIC large-category classifications. In this section, we use the GPT-L English specification, which later yields the highest alpha in the lead-lag analysis reported in Section 4.2.

Figure 2 presents the case of SK hynix. Firms such as SAMSUNG ELECTRONICS and SAMSUNG ELECTRO-MECHANICS, which are classified in the same KSIC large category (Electrical and Electronic) and are also treated as comparable firms in FnGuide, maintain consistently high similarity over time. By contrast, GAON CABLE, KYUNG DONG NAVIEN, and DAWON CABLE exhibit persistently low similarity despite belonging to the same KSIC category. This pattern suggests that the proposed network distinguishes firms with materially different business profiles even when they are grouped together under the conventional industry classification. UNIQUEST also exhibits consistently high cosine similarity with SK hynix despite being assigned to a different KSIC category (Distribution). This result likely reflects the fact that, although KSIC classifies UNIQUEST as a distributor, publicly available firm descriptions characterize it as a semiconductor-related solution and distribution company, making its business profile closer

〈Figure 2〉 SK hynix



This figure plots the time-series variation in cosine similarity between SK hynix and selected firms over the period 2010–2024.

to the semiconductor value chain than its KSIC label would suggest. In contrast, affiliates such as SK Innovation receive consistently low similarity, suggesting that the network reflects business activities rather than group affiliation.

Figure 3 provides a parallel illustration using AMOREPACIFIC as the focal firm. Although AMOREPACIFIC is classified under the KSIC large category of Chemicals, its business descriptions show consistently high similarity with cosmetics-related firms such as TONYMOLY, LG H&H, COSMAX, and HANKOOK COSMETICS. By contrast, firms such as LOTTE CHEMICAL, S-Oil, and KUMHO TIRE, which are also grouped into the same KSIC large category, remain at much lower similarity levels throughout the sample period, as their core business activities differ substantially from cosmetics-related operations. Taken together, these examples suggest that the LLM-based business network captures stable business proximity more effectively than broad industry labels alone.

〈Figure 3〉 AMOREPACIFIC

This figure plots the time-series variation in cosine similarity between AMOREPACIFIC and selected firms over the period 2010–2024.

4.2 Lead-lag effect

Prior studies show that lead-lag effects in stock returns are stronger among economically related firms, including in the Korean market. Motivated by this literature, we examine whether the proposed business network captures return-relevant information more effectively than the conventional KSIC classification. Following Breitung and Müller (2025), we estimate intercept alphas by regressing long-short returns constructed from business-network signals on the baseline 7 factor model described in Section 2.2.

The lead-lag analysis is designed primarily as an economic validation test of the proposed business networks rather than as a direct evaluation of an implementable trading strategy. If a business network captures economically meaningful peer relationships, returns of closely related peer firms should contain information about the focal firm's subsequent returns beyond standard risk factors and conventional industry classifications. Accordingly, we interpret the resulting alphas as evidence on the informational content of the network, not as net trading profits. The reported alphas are gross monthly alphas before transaction costs. To partly address concerns that the results may be driven by small or illiquid firms, we report both equal-weighted and value-weighted specifications throughout the analysis.

Using the peer definitions introduced in Section 3.3, we construct monthly rebalanced calendar-time portfolios from July of year t+1 to June of year t+2, consistent with the timing convention in Fama and French (1993). In each month t, firms are ranked according to the return signal implied by their peer firms' returns in month t-1. We consider two weighting schemes for this peer-return signal. The signal is defined as the average return of the focal firm's peers, computed as a simple average in the equal-weighted specification and as a market-capitalization-weighted average in the value-weighted specification. In both cases, the final long-short factor is formed by taking equal positions in firms in the top 10% (or 20%) of the ranking and short positions in firms in the bottom 10% (or 20%).

Prior to examining the lead-lag results, it is useful to assess how closely the proposed LLM-based business networks overlap with the conventional KSIC classification. Using the peer definitions introduced in Section 3.3, peer firms are identified each year based on annual cosine-similarity cutoffs corresponding to the top 1%, 2%, 5%, and 10% of firm-pair similarities. We then measure the degree of overlap between the LLM-based networks and the KSIC benchmark using the Jaccard similarity index. Table 3 reports the average overlap across sample years by the output language of the LLM-generated business descriptions and by peer-selection threshold. The overlap remains partial across all specifications, suggesting that the proposed networks capture a dimension of firm-level relatedness that is not fully reflected in the conventional KSIC classification. This comparison provides a useful benchmark for interpreting the subsequent asset-pricing results.

〈Table 3〉 Overlap with KSIC

This table reports the average Jaccard similarity between the LLM-based business networks and the KSIC-based network across sample years. The overlap is reported separately by the output language of the LLM-generated business descriptions and by the peer-selection threshold used to define network links, where peer firms are identified from the top 1%, 2%, 5%, and 10% of annual firm-pair cosine similarities. Higher values indicate greater overlap with the conventional KSIC classification.

Language	Peer Firms Selection	BERT	FinBERT	Gemini	GPT-L
English	Top 1%	5.56%	6.48%	12.01%	12.88%
	Top 2%	7.18%	8.70%	15.51%	16.11%
	Top 5%	9.68%	11.92%	18.99%	19.51%
	Top 10%	11.02%	13.72%	19.06%	19.39%
Korean	Top 1%	1.93%	5.43%	12.41%	10.21%
	Top 2%	2.66%	6.99%	16.01%	13.17%
	Top 5%	4.01%	9.48%	19.31%	16.58%
	Top 10%	5.17%	10.91%	19.02%	17.00%

For comparison, we also construct benchmark factors using the traditional KSIC large-category classification, where peer firms are defined by membership in the same KSIC group. As reported in Table 4, the KSIC-based benchmark does not generate statistically significant alpha under either weighting scheme. Although the value-weighted KSIC factor performs somewhat better than its equal-weighted counterpart, with alphas of 34.65 bps and 13.84 bps under the 10% and 20% portfolio-formation rules, respectively, these estimates remain statistically insignificant. These results suggest that conventional industry classifications alone provide limited information for capturing the lead-lag structure in returns.

〈Table 4〉 Lead-lag effect for KSIC Network

This table reports the alphas of KSIC-based lead-lag factors under equal-weighted and value-weighted specifications. The results indicate that the KSIC benchmark does not generate statistically significant alpha under either portfolio-formation rule. t-values are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Weight	Factor Formation	KSIC Factor
Equal-weighted	Top 10% - Bottom 10%	-0.0087 (-0.027)
	Top 20% - Bottom 20%	0.0709 (0.358)
Value-weighted	Top 10% - Bottom 10%	0.3465 (1.560)
	Top 20% - Bottom 20%	0.1384 (0.788)

Because the empirical design compares multiple output languages, embedding models, peer-selection thresholds, and portfolio-formation rules, we do not interpret the results as identifying a single optimal specification. Instead, our focus is on whether LLM-based business networks, as a class of text-based peer definitions, contain return-relevant information beyond the KSIC benchmark and standard risk factors. We therefore emphasize patterns that are consistent across related specifications and report the full set of results, rather than relying on a single selected model.

Table 5 reports the results for the equal-weighted LLM-based business-network factors. Across a wide range of peer-selection thresholds and portfolio-formation rules, the proposed factors generate positive and often statistically significant alphas. Relative to the KSIC benchmark in Table 4, the LLM-based factors provide substantially stronger evidence of return predictability. For the English-output specification (Panel A), the highest alpha is 53.53 bps (t = 1.93), achieved by GPT-L at the top 5% peer threshold with the top 10% minus bottom 10% formation rule. For the Korean-output specification (Panel B), the strongest result is 50.53 bps (t = 2.94), achieved by Gemini at the top 1% threshold with the top 20% minus bottom 20% rule. Overall, the equal-weighted results indicate that the business-network factors contain return-relevant information beyond standard factor exposures and outperform the KSIC-based benchmark. At the same time, the results suggest a relative advantage of the English-output specification in terms of the breadth and stability of significant alphas, while Korean-output specifications also produce strong and economically meaningful alphas in selected cases.

〈Table 6〉 Lead-lag effect for LLM-based Business Network (Value-weighted)

This table reports the alphas of lead-lag factors constructed from the LLM-based business network under the value-weighted specification. t-values are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Peer Firms	Factor Formation	BERT	FinBERT	Gemini	GPT-L
Panel A: English					
Top 1%	Top 10% - Bottom 10%	0.4026** (2.128)	0.3821* (1.823)	0.0708 (0.268)	0.3341 (1.515)
	Top 20% - Bottom 20%	0.3164** (2.132)	0.3662** (2.536)	0.3192* (1.739)	0.3544** (2.147)
Top 2%	Top 10% - Bottom 10%	0.2361 (1.126)	0.4599** (2.467)	0.2340 (1.136)	0.4466** (2.088)
	Top 20% - Bottom 20%	0.1663 (1.182)	0.3660** (2.437)	0.3429** (2.128)	0.3156* (1.918)
Top 5%	Top 10% - Bottom 10%	0.2270 (0.970)	0.5728** (2.353)	0.3819 (1.347)	0.2546 (1.017)
	Top 20% - Bottom 20%	0.0661 (0.367)	0.3369* (1.816)	0.4164** (2.035)	0.3053* (1.667)
Top 10%	Top 10% - Bottom 10%	0.3547 (1.620)	0.6967*** (2.857)	0.3871 (1.532)	0.2669 (1.077)
	Top 20% - Bottom 20%	0.1593 (0.827)	0.3313 (1.601)	0.4118* (1.817)	0.3146 (1.509)
Panel B: Korean					
Top 1%	Top 10% - Bottom 10%	-0.0140 (-0.082)	-0.2969 (-1.565)	0.5117** (2.127)	0.4563* (1.917)
	Top 20% - Bottom 20%	0.1036 (0.885)	-0.1543 (-1.122)	0.5188*** (2.666)	0.3872** (1.990)
Top 2%	Top 10% - Bottom 10%	-0.0826 (-0.424)	0.1968 (0.898)	0.4576* (1.827)	0.3625* (1.821)
	Top 20% - Bottom 20%	-0.0321 (-0.227)	0.2655 (1.443)	0.4214** (2.190)	0.3861** (2.550)
Top 5%	Top 10% - Bottom 10%	0.1164 (0.642)	0.4230* (1.783)	0.3580 (1.386)	0.3632 (1.362)
	Top 20% - Bottom 20%	0.1116 (0.813)	0.2678 (1.466)	0.3785* (1.756)	0.0896 (0.486)
Top 10%	Top 10% - Bottom 10%	0.1180 (0.638)	0.6999*** (2.927)	0.5395** (2.206)	0.3662 (1.504)
	Top 20% - Bottom 20%	0.0908 (0.567)	0.4342** (2.204)	0.4678** (2.241)	0.2936* (1.688)

Tables 5 and 6 indicate that the output language of LLM-generated business descriptions matters for the resulting business-network signal. English-output specifications tend to produce more stable and broadly significant alphas across embedding models and portfolio-formation rules, whereas Korean-output specifications also generate economically large alphas in selected cases, particularly under value-weighted schemes. One possible interpretation is that information from economically related large firms diffuses more strongly to the returns of connected firms under value-weighted specifications, consistent with Lo and MacKinlay (1990) and Hou (2007). Thus, the evidence should not be interpreted as showing that English summaries uniformly dominate Korean summaries. Rather, the results suggest that the output language interacts with the embedding architecture and the portfolio-construction rule in shaping the similarity structure and its return-predictive content.

One possible interpretation is that English summaries generated from Korean disclosures may benefit from greater standardization in financial and business terminology, as many large-scale embedding models are trained on extensive English-language corpora. This standardization can make semantically related firms appear closer in the embedding space even when the original Korean disclosures differ in wording or reporting style. At the same time, the strong performance of several Korean-output specifications indicates that the original-language representation also preserves economically meaningful business information. The relevant empirical implication is therefore not that one language is universally superior, but that output-language choice is a substantive modeling decision in LLM-based financial text analysis.

Finally, Table 7 and Table 8 re-examine the specifications that yielded statistically significant alphas using an eight-factor benchmark that augments the seven-factor model with the KSIC factor. Since the full set of eight-factor results is reported in Appendix B, Tables 7 and 8 provide compact summaries by reporting one representative significant specification for each embedding model and output-language pair under the equal-weighted and value-weighted schemes, respectively. These tables are intended to summarize whether the main findings survive the inclusion of the KSIC factor, rather than to identify an ex-post optimal trading rule.

Despite controlling for the KSIC factor in addition to the standard risk factors, many specifications remain positive and statistically significant. In particular, GPT-L remains significant in the equal-weighted English-output specification, while FinBERT continues to show especially strong performance in the value-weighted results for both output-language settings. These results suggest that the strongest signals captured by the proposed business-network factors are not fully subsumed by either the standard risk factors or the KSIC benchmark. At the same time, the selected-specification format of Tables 7 and 8 suggests that differences across output languages should be interpreted with some caution. The more appropriate conclusion is that the English-output specification generally shows more stable performance across specifications, while Korean-output specifications also retain substantial predictive content in selected cases under the expanded

benchmark. These results suggest that the proposed business-network factors contain information beyond that captured by standard risk factors and the KSIC benchmark.

〈Table 7〉 Lead-lag effect with 7 factor and KSIC factor (Equal-weighted)

This table reports the alphas of equal-weighted LLM-based lead-lag factors using the eight-factor benchmark that includes the KSIC factor, showing that many specifications remain statistically significant after controlling for the KSIC benchmark. t-values are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Language	Model	Peer Firms Selection	Factor Formation	Alpha
English	BERT	Top 1%	Top 20% - Bottom 20%	0.4059*** (3.00)
	FinBERT	Top 5%	Top 10% - Bottom 10%	0.5305** (2.29)
	Gemini	Top 10%	Top 10% - Bottom 10%	0.5082** (2.02)
	GPT-L	Top 5%	Top 10% - Bottom 10%	0.5067** (2.02)
Korean	BERT	Top 5%	Top 10% - Bottom 10%	0.3873* (1.73)
	FinBERT	Top 10%	Top 10% - Bottom 10%	0.4079* (1.87)
	Gemini	Top 1%	Top 20% - Bottom 20%	0.4914*** (2.77)
	GPT-L	Top 1%	Top 10% - Bottom 10%	0.3353* (1.81)

〈Table 8〉 Lead-lag effect with 7 factor and KSIC factor (Value-weighted)

This table reports the alphas of value-weighted LLM-based lead-lag factors using the eight-factor benchmark that includes the KSIC factor, showing that many specifications remain statistically significant after controlling for the KSIC benchmark. t-values are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Language	Model	Peer Firms Selection	Factor Formation	Alpha
English	BERT	Top 1%	Top 10% - Bottom 10%	0.3498* (1.841)
	FinBERT	Top 10%	Top 10% - Bottom 10%	0.6000** (2.508)
	Gemini	Top 5%	Top 20% - Bottom 20%	0.3487* (1.868)
	GPT-L	Top 2%	Top 10% - Bottom 10%	0.3480* (1.667)

Korean	BERT	-	-	-
	FinBERT	Top 10%	Top 10% - Bottom 10%	0.5692** (2.392)
	Gemini	Top 1%	Top 20% - Bottom 20%	0.4604*** (2.719)
	GPT-L	Top 2%	Top 20% - Bottom 20%	0.3354** (2.514)

[5] Conclusion

This study examines whether the output language of LLM-generated business descriptions affects the construction and empirical performance of text-based business networks in a non-English disclosure environment. Using annual reports of KOSPI-listed firms from 2010 to 2024, we generate Korean and English business summaries from the same underlying Korean disclosures and construct firm-level business networks using several embedding models. We then evaluate whether these networks capture economically meaningful peer relationships through temporal-stability analyses, overlap comparisons with KSIC classifications, and peer-based lead-lag return tests.

The results show that the proposed LLM-based business networks identify firm-level economic relatedness that is not fully captured by conventional industry classifications. The networks exhibit stable similarity patterns over time and overlap only partially with KSIC-based peer groups. In the asset-pricing tests, KSIC-based lead-lag factors do not generate statistically significant alphas, whereas several LLM-based business-network factors produce positive and statistically significant alphas under both equal-weighted and value-weighted specifications. These results remain statistically and economically meaningful even after controlling for standard risk factors and a KSIC-based benchmark factor, suggesting that the proposed networks contain information beyond traditional industry classifications.

The findings also indicate that the output language of LLM-generated business descriptions is an economically relevant modeling choice. English summaries generated from Korean disclosures tend to produce more stable and broadly significant lead-lag performance across specifications, while Korean summaries also generate strong alphas in selected combinations of models, thresholds, and weighting schemes. Therefore, the results should not be interpreted as showing that English representations uniformly dominate Korean representations. Rather, they suggest that output language interacts with embedding architecture and portfolio-construction choices in shaping the resulting similarity structure and its return-predictive content.

Several limitations should be noted. First, al-

though the analysis uses comparable prompt structures and generation settings, it does not fully separate pure language effects from LLM summarization effects and embedding-model differences. Second, while the masking and verification procedures reduce look-ahead concerns, residual identifiability cannot be ruled out completely. Third, the lead-lag tests are intended to validate the economic content of the proposed networks rather than to establish a directly implementable trading strategy. Future research could incorporate transaction costs, liquidity constraints, short-sale restrictions, and more formal multiple-testing adjustments. Future work could also compare LLM-generated summaries with direct machine translations, raw-text embeddings, and English-language source disclosures as they become more widely available in the Korean market.

〈Table 5〉 Lead-lag effect for LLM-based Business Network (Equal-weighted)

This table reports the alphas of lead-lag factors constructed from the LLM-based business network under the equal-weighted specification. t-values are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Peer Firms	Factor Formation	BERT	FinBERT	Gemini	GPT-L
Panel A: English					
Top 1%	Top 10% - Bottom 10%	0.3924*** (2.59)	0.2899 (1.42)	0.1556 (0.65)	0.3652 (1.57)
	Top 20% - Bottom 20%	0.4145*** (3.08)	0.3904*** (2.87)	0.0971 (0.51)	0.2517 (1.49)
Top 2%	Top 10% - Bottom 10%	0.3397** (2.17)	0.2561 (1.22)	-0.1057 (-0.45)	0.4705* (1.81)
	Top 20% - Bottom 20%	0.2772** (1.97)	0.3546** (2.42)	0.1893 (1.09)	0.3774** (2.17)
Top 5%	Top 10% - Bottom 10%	0.0143 (0.07)	0.5267** (1.99)	0.3932* (1.67)	0.5353* (1.93)
	Top 20% - Bottom 20%	0.0408 (0.21)	0.4624** (2.52)	0.3197 (1.53)	0.4596** (2.17)
Top 10%	Top 10% - Bottom 10%	0.1713 (0.83)	0.2389 (0.96)	0.5041* (1.85)	0.4010 (1.48)
	Top 20% - Bottom 20%	0.2499 (1.34)	0.4331** (2.20)	0.3438* (1.71)	0.3594 (1.53)
Panel B: Korean					
Top 1%	Top 10% - Bottom 10%	0.3124 (1.43)	-0.0284 (-0.13)	0.3340 (1.61)	0.3328* (1.69)
	Top 20% - Bottom 20%	0.0652 (0.49)	-0.0374 (-0.26)	0.5053** (2.94)	0.2563 (1.54)
Top 2%	Top 10% - Bottom 10%	0.2671 (1.21)	0.1534 (0.75)	-0.0643 (-0.30)	0.3858* (1.79)

	Top 20% - Bottom 20%	0.2008 (1.40)	0.1862 (0.99)	0.2068 (1.15)	0.2255 (1.58)
Top 5%	Top 10% - Bottom 10%	0.3869* (1.69)	0.3132 (1.35)	0.1942 (0.67)	0.0660 (0.25)
	Top 20% - Bottom 20%	0.0618 (0.36)	0.3586* (1.89)	0.3222 (1.37)	0.2359 (1.02)
Top 10%	Top 10% - Bottom 10%	0.1025 (0.45)	0.4056* (1.70)	0.1945 (0.64)	0.3095 (1.05)
	Top 20% - Bottom 20%	0.1136 (0.62)	0.2731 (1.29)	0.4385* (1.90)	0.1857 (0.74)

Table 6 presents the corresponding value-weighted results, where the peer-return signal places greater weight on larger peer firms. The value-weighted specification also produces strong and statistically significant alphas, and in several cases the performance exceeds that of the equal-weighted specification. In the English-output specification (Panel A), the highest alpha rises to 69.67 bps (t = 2.86), achieved by FinBERT at the top 10% peer threshold with the top 10% minus bottom 10% rule. In the Korean-output specification (Panel B), the strongest result is 69.99 bps (t = 2.93), also achieved by FinBERT at the top 10% threshold with the top 10% minus bottom 10% rule. These findings indicate that the relative advantage of the English-output specification is not uniform across all cases. Nevertheless, the English-output specification still tends to deliver more stable and broadly significant performance across specifications, whereas Korean-output specifications also generate high alphas under particular model-threshold-weighting combinations. One possible interpretation is that English summaries generated from Korean disclosures may, in some cases, align more closely with embedding spaces shaped by large English or multilingual corpora and may also present financial terminology in a more standardized form. At the same time, the strong performance of several Korean-output specifications suggests that such an advantage is conditional rather than universal.

References

Breitung, C. and S. Müller, 2025, Global business networks, Journal of Financial Economics, Vol. 166, pp. 104007.

Cen, L., K. Chan, S. Dasgupta, and N. Gao, 2013, When the tail wags the dog: Industry leaders, limited attention, and spurious cross-industry information diffusion, Management Science, Vol. 59, No. 11, pp. 2566-2585.

Fama, E. F. and K. R. French, 1993, Common risk factors in the returns on stocks and bonds, Journal of Financial Economics, Vol. 33, No. 1, pp. 3-56.

Fama, E. F. and K. R. French, 2015, A five-factor asset pricing model, Journal of Financial Economics, Vol. 116, No. 1, pp. 1-22.

Glasserman, P. and C. Lin, 2023, Assessing look-ahead bias in stock return predictions generated by GPT sentiment analysis, Journal of Financial Data Science, Vol. 6, No. 1, pp. 25-42.

Han, M., D. H. Lee, and H. G. Kang, 2020, Market anomalies in the Korean stock market, Journal of Derivatives and Quantitative Studies, Vol. 28, No. 2, pp. 3-50.

Hoberg, G. and G. Phillips, 2010, Product market synergies and competition in mergers and acquisitions: A text-based analysis, The Review of Financial Studies, Vol. 23, No. 10, pp. 3773-3811.

Hoberg, G. and G. Phillips, 2016, Text-based network industries and endogenous product differentiation, Journal of Political Economy, Vol. 124, No. 5, pp. 1423-1465.

Hou, K., 2007, Industry information diffusion and the lead-lag effect in stock returns, The Review of Financial Studies, Vol. 20, No. 4, pp. 1113-1138.

Jegadeesh, N., 1990, Evidence of predictable behavior of security returns, The Journal of Finance, Vol. 45, No. 3, pp. 881-898.

Jegadeesh, N. and S. Titman, 1993, Returns to buying winners and selling losers: Implications for stock market efficiency, The Journal of Finance, Vol. 48, No. 1, pp. 65-91.

Kang, S. H. and S. M. Yoon, 2012, Asymmetric information transmission among different size stock indexes in the Korean stock market, Korean Journal of Business Administration, Vol. 25, No. 6, pp. 2853-2869.

Lee, J. et al., 2025, Gemini embedding: Generalizable embeddings from gemini, arXiv preprint arXiv:2503.07891.

Lee, S. and J.-G. Park, 2021, A comparative study of stock & cryptocurrency market prediction using internet search volume, Asset Management Review, Vol. 9, No. 2, pp. 41-61.

Lo, A. W. and A. C. MacKinlay, 1990, When are contrarian profits due to stock market overreaction?, The Review of Financial Studies, Vol. 3, No. 2, pp. 175-205.

Park, S. Y. and J. H. Heo, 2020, The lead-lag relationship between consumer sentiment index and large, middle and small capitalization indexes of KOSPI, Journal of Corporation and Innovation, Vol. 43, No. 4, pp. 111-129.

Park, Y. G. and S. Y. Jang, 2003, Trading volume and lead-lag effect in Korean stock market, Korean Journal of Financial Studies, Vol. 32, No. 2, pp. 105-139

Woo, Min-cheol and Jee-hyun Kim, 2017, Can the cyber-information influence the stock price? : Using an artificial intelligent(AI) algorithm methodology, Asset Management Review, Vol. 5, No. 2, pp. 40-55.

Yang, Y., M. C. S. Uy, and A. Huang, 2020, FinBERT: A pretrained language model for financial communications, arXiv preprint arXiv:2006.08097.

Appendix

A. Prompt Template for Business Description Generation

A.1 English version

[Role]

You are a top-tier financial analyst. Your task is to analyze corporate documents provided in Korean and synthesize them into a professional English summary.

[Objective]

Comprehensively analyze the two provided Korean text segments to write a summary in English that clearly explains the company's core business model.

- Source 1: 'Business Overview'
- Source 2: 'Management's Discussion and Analysis'

[Instructions]

✓ Write in English.
✓ Do not exceed 512 tokens (Target Token Count: Between 400 and 512 output tokens)
✓ Write in a paragraph format using a professional and objective tone.
✓ If possible, specify the business segment that accounts for the largest share of total revenue.
✓ Seamlessly integrate descriptions of business segments (qualitative information) with numerical data from tables (quantitative information).
✓ Use only the provided text and tables as the sole basis for your analysis. Do not use any prior knowledge or external information regarding the company.

A.2 Korean version

[역할]

당신은 주어진 기업의 설명을 요약하는 최고의 금융 애널리스트입니다. 당신의 분석은 논리적이고, 객관적이며, 반드시 주어진 데이터에만 근거해야 합니다.

[목표]

아래 주어지는 두 가지 텍스트 데이터를 종합적으로 분석하여, 해당 기업의 '핵심 비즈니스 모델'을 명확히 설명하는 요약문을 작성합니다.

- 첫 번째 텍스트: '사업의 개요'
- 두 번째 텍스트: '이사의 경영진단 및 분석의견'

[지시사항]

✓ 한글로 512토큰을 넘어가지 않도록 작성하세요. (400토큰 이상 512토큰 이하로 작성)
✓ 문단 형태, 전문가적이고 객관적인 톤으로 작성하세요.
✓ 가능하다면 전체 매출에서 차지하는 비중이 가장 높은 사업 부분을 명시하세요.
✓ 사업 부문에 대한 설명(정성적 정보)과 테이블의 수치 데이터(정량적 정보)를 자연스럽게 연결하여 서술하세요.
✓ 주어진 텍스트와 테이블의 내용만을 분석의 유일한 근거로 사용하고, 당신이 사전에 알고 있는 해당 기업에 대한 정보나 외부 지식을 절대 사용해서는 안 됩니다.

B. Lead-lag effect with 7 factor and KSIC factor

B.1 Lead-lag effect with 7 factor and KSIC factor (Equal-weighted)

This table reports the alphas of equal-weighted LLM-based lead-lag factors using the eight-factor benchmark that includes the KSIC factor, showing that many specifications remain statistically significant after controlling for the KSIC benchmark. t-values are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Peer Firms Selection	Factor Formation	BERT	FinBERT	Gemini	GPT-L
Panel A: English					
Top 1%	Top 10% - Bottom 10%	0.3941*** (2.61)	-	-	-
	Top 20% - Bottom 20%	0.4059*** (3.00)	0.3713*** (2.89)	-	-
Top 2%	Top 10% - Bottom 10%	0.3406** (2.04)	-	-	0.4680* (1.78)
	Top 20% - Bottom 20%	0.2660* (1.86)	0.3324** (2.46)	-	0.2609 (1.54)
Top 5%	Top 10% - Bottom 10%	-	0.5305** (2.29)	0.3967* (1.66)	0.5067** (2.02)
	Top 20% - Bottom 20%	-	0.4297*** (2.63)	-	0.4022* (1.80)
Top 10%	Top 10% - Bottom 10%	-	-	0.5082** (2.02)	-
	Top 20% - Bottom 20%	-	0.4054** (2.16)	0.3130 (1.57)	-
Panel B: Korean					
Top 1%	Top 10% - Bottom 10%	-	-	-	0.3353* (1.81)
	Top 20% - Bottom 20%	-	-	0.4914*** (2.77)	-
Top 2%	Top 10% - Bottom 10%	-	-	-	0.2108 (0.82)
	Top 20% - Bottom 20%	-	-	-	-
Top 5%	Top 10% - Bottom 10%	0.3873* (1.73)	-	-	-
	Top 20% - Bottom 20%	-	0.3334* (1.82)	-	-
Top 10%	Top 10% - Bottom 10%	-	0.4079* (1.87)	-	-
	Top 20% - Bottom 20%	-	-	0.4083* (1.91)	-

B.2 Lead-lag effect with 7 factor and KSIC factor (Value-weighted)

This table reports the alphas of value-weighted LLM-based lead-lag factors using the eight-factor benchmark that includes the KSIC factor, showing that many specifications remain statistically significant after controlling for the KSIC benchmark. t-values are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Peer Firms	Factor Formation	BERT	FinBERT	Gemini	GPT-L
Panel A: English					
Top 1%	Top 10% - Bottom 10%	0.3498* (1.841)	0.2910 (1.341)	-	-
	Top 20% - Bottom 20%	0.2769** (2.005)	0.3224** (2.454)	0.2812 (1.598)	0.2979* (1.940)
Top 2%	Top 10% - Bottom 10%	-	0.3600* (1.802)	-	0.3480* (1.667)
	Top 20% - Bottom 20%	-	0.3205** (2.133)	0.2894** (2.032)	0.2548* (1.747)
Top 5%	Top 10% - Bottom 10%	-	0.4358* (1.794)	-	-
	Top 20% - Bottom 20%	-	0.2778 (1.562)	0.3487* (1.868)	0.2362 (1.492)
Top 10%	Top 10% - Bottom 10%	-	0.6000** (2.508)	-	-
	Top 20% - Bottom 20%	-	-	0.3449* (1.702)	-
Panel B: Korean					
Top 1%	Top 10% - Bottom 10%	-	-	0.4156* (1.740)	0.3430 (1.591)
	Top 20% - Bottom 20%	-	-	0.4604*** (2.719)	0.3345** (2.056)
Top 2%	Top 10% - Bottom 10%	-	-	0.3239 (1.413)	0.2781 (1.469)
	Top 20% - Bottom 20%	-	-	0.3607** (2.186)	0.3354** (2.514)
Top 5%	Top 10% - Bottom 10%	-	0.3102 (1.288)	-	-
	Top 20% - Bottom 20%	-	-	0.3142 (1.580)	-
Top 10%	Top 10% - Bottom 10%	-	0.5692** (2.392)	0.4310* (1.843)	-
	Top 20% - Bottom 20%	-	0.3742** (1.992)	0.4100** (2.111)	0.2432 (1.551)

LLM 기반 비즈니스 네트워크에서 언어 선택은 중요한가? 한국 기업공시를 이용한 실증분석*

이선아** (성균관대학교)
임병화*** (성균관대학교)

초 록

본 연구는 KOSPI 상장기업의 사업보고서를 활용하여 대규모 언어모형(LLM) 기반 동적 비즈니스 네트워크를 구축하고, 공시 정보를 표현하는 언어가 네트워크의 경제적 정보성에 미치는 영향을 분석하였다. 2010년부터 2024년까지 사업보고서의 '사업의 내용'과 '이사의 경영진단 및 분석의견'을 추출한 뒤, 동일한 원천 공시자료를 바탕으로 한글과 영어의 표준화된 사업 설명문을 각각 생성하였다. 사전학습 정보에 따른 look-ahead bias를 완화하기 위해 기업명과 제품명은 LLM 생성 이전에 익명화하였다. 이후 다양한 임베딩 모형을 이용하여 기업 간 코사인 유사도를 측정하고, 연도별 비즈니스 네트워크를 구성하였다. 분석 결과, 제안한 네트워크는 기존 한국표준산업분류(KSIC)가 포착하지 못하는 기업 수준의 경제적 유사성을 식별하였다. 또한 동일한 공시자료라도 한글과 영어 설명문은 상이한 유사도 구조와 선후행 수익률 신호를 생성하였다. 자산가격결정 분석에서는 네트워크 기반 동종기업 수익률 요인이 표준 위험요인과 KSIC 기반 요인을 통제한 이후에도 유의한 알파를 나타냈다. 특히 영어 기반 설명문은 여러 주요 설정에서 보다 안정적이고 폭넓게 유의한 선후행 성과를 보였으며, 한글 설명문도 일부 설정에서 의미 있는 예측 정보를 제공하였다. 이는 LLM 기반 금융 텍스트 분석에서 언어 선택이 단순한 전처리 절차가 아니라 중요한 모형 설계 선택임을 시사한다.

* 본 논문의 심사과정에서 유익한 논평을 해주신 심사위원들께 감사드립니다. 이 논문은 주저자의 석사학위논문의 일부를 수정·보완한 것입니다. 본 연구는 2026년도 교육부 및 서울특별시 재원으로 서울 RISE센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE)의 결과입니다(2026-RISE-01-018-04).
주제어: 산업분류, 비즈니스 네트워크, 대규모 언어모형, 텍스트 분석, 선후행 효과
JEL 분류기호: G12, G14, C55, L16
** 주저자. 성균관대학교 핀테크융합전공, e-mail: dltjsdk0321@gmail.com
*** 교신저자. 성균관대학교 핀테크융합전공, 경영대학, e-mail: limbh@skku.edu