Journal of High School Science

Utilizing Large Language Models for text-based Industry classification

Offutt J

Submitted: January 28, 2025, Revised: version 1, April 10, 2025, version 2, May 11, 2025
Accepted: May 12, 2025

Abstract

This study develops a novel, dynamic industry classification system, rooted in Artificial Intelligence (AI), by using Large Language Model (LLM) technology to analyze and compare firms' product descriptions as found in Securities and Exchange Commission (SEC) 10-Q and 10-K filings. Unlike traditional static classification systems such as the Standard Industrial Classification (SIC) or the North American Industry Classification System (NAICS), the proposed method dynamically quantifies the degree of competition and customer-supplier relationships between firms. It utilized a 210x210 similarity matrix to compile the relationship scores as a starting point for further analysis. This enhanced metric strengthens the literature and aids in the identification of portfolio correlations, providing nuanced firm-to-firm insights that other methodologies have not fully captured. In turn, this assists investors in risk management and provides insights into behavioral finance by highlighting how news perception affects market dynamics. It also has potential implications on merger and acquisition strategy, supply chain analysis, and policy making. The methodology employs Ordinary Least Squares (OLS) regression and pairwise correlation analysis to evaluate the efficacy of the LLM measurement against the SIC and NAICS codes. The LLM outperformed the other methods across most models. In the few cases where it did not, the models had low observation counts, lower R^2 values, and weak F-statistics. These findings indicate that the utilization of LLMs and AI as an industry classification tool is plausible and superior to the customary past measures of competitors and especially of customer-supplier identification for the majority of industry code granularities.

Keywords

Industry classification, Artificial Intelligence, Large Language Model, Industry cluster, Economic linkage, Regression analysis, SEC filings, North American Industry Classification System, ChatGPT API

James Offutt, The Hotchkiss School, 11 Interlaken Rd, Lakeville, CT 06039, USA. Jamesoffutt3@icloud.com

https://doi.org/10.64336/001c.138087

Introduction

Classifying firms' industry and the economic linkages within related industries is central to the studies of industrial organization, portfolio theory, and arbitrage opportunities. This research examines whether a Large Language Model (LLM) driven text-based methodology for industry classification offers a superior measurement of industry relationships and identification of linked firms. This paper diverged from existing industry metrics such as SIC, instead associating firms by analyzing their product descriptions from SEC quarterly and yearly filings using cutting-edge LLM technology. The method not only refined the understanding of firms' economic roles as competitors and customer-suppliers but also enhanced market participants' and academics' understanding of market behavior. Having enabled more accurate identification of firm linkages, the proposed methodology aids in uncovering otherwise hidden correlations within portfolio holdings, thereby facilitating more effective risk management. Additionally, this study contributes to the field of behavioral finance by exploring how investor perceptions and reactions to news about interconnected companies can influence market trends. These insights can help investors build strategies that respond to market psychology, while also improving market efficiency.

If an investment strategy is intended to be industry-neutral but was constructed using SIC codes, it might consider a long position in Pfizer and short Eli Lilly to balance industry exposure. However, this paper's refined measure reveals a discrepancy: Eli Lilly has a strong presence in traditional pharmaceutical markets where Pfizer is less dominant. Yet, Pfizer has been significantly involved in the COVID-19 sector, unlike Eli Lilly. Implementing this strategy during a period where COVID-19 is less relevant and other attributes, such as weight loss, are more prominent, could expose investors to unexpected industry risks. This example underscores the potential value of a more nuanced industry classification system, which is crucial not only for investment strategies but also for academic research. Finance and accounting studies frequently discuss industry and information spillovers, predicated on the assumption that companies within the same industry will exhibit correlated financial behaviors such as earnings and expenses. A more precise measure of industry, as proposed by this methodology, is essential to accurately assess these spillovers, with implications spanning earnings, litigation, and product recalls. The enhanced classification system constructed in this paper promises to contribute significantly to both investment management and academic literature in finance and accounting.

This study relates to the existing body of literature as we address similar concerns with the identification of industry categories. Much like this paper, prior literature has recognized the inefficiencies and risks of static industry classifications such as SIC codes and the need to ascertain a new, more dynamic and reliable measure of industry categorization. In particular, the paper confronts the same endeavor of identifying a new industry measurement methodology while using the previously examined concept of text-based analysis. However, this approach does this by utilizing more up-to-date technology that relies on LLM technology. Papers such as those by Guenther and Rosman, Jacobs and O'Neill, and

Marozzi underscore the limitations associated with the methodology behind SIC codes, while also encouraging further exploration of alternative measurements (1-3). In particular, this segment of the literature advocates for using firm-specific data, containing characteristics of firms, rather than depending on self-assigned industry classifications such as those used in SIC codes. Research such as that by Hoberg and Phillips, Papagiannidis et al., and Murdock et al. attempt to expand upon this constraint within the industry categorization literature by utilizing text-based approaches and firm data (4-6). These papers rely on a range of approaches; from data in SEC filings, to vast sets of internet-mined data.

The focus and findings of this prior body of research raises a series of considerations: the accuracy of SIC codes, the lack of nuance within industry measures, and the reliability of more dynamic measures. These concerns created the necessity for this study. The SIC-focused portion of the existing literature has outlined key concerns such as: companies' SIC codes that vary depending on the reporting source (Compustat® versus CRSP), and the static nature of these codes meaning that they are unable to match the current competitive sector of firms. Furthermore, the existing body of literature highlights the lack of nuance captured by previous industry measures such as SIC. Since SIC codes are not based on firm-specific data and are static, they usually only capture a firm's primary industry, missing more nuanced firm-to-firm competition. This feature limits industry classification codes to solely encapsulating the big-picture of firms instead of capturing the nuanced details that could call attention to important firm-to-firm relationships and less recognizable sectors of businesses.

The body of literature and academic awareness around the short-comings of existing industry codes has continued to grow at the same time that the economic rationale for classification refinement has developed and the perceived value of dynamic industry measures has increased. Whether it be for the increased adoption of higher frequency and algorithm-based trading or for more accurate and elastic measures of industry for the purpose of academic pursuits, the need for dynamic measures has never been greater. However, as more dynamic classification studies have emerged, concerns around data selection, standardization, the overweighting of certain changes, and the lack of user-friendly interpretation have become prominent. This is the genesis of the research question. This paper aimed to address these concerns by [1] relying on SEC filings for robust, reliable, and periodically reported data, [2] using a series of processes and thresholds to maximize standardization, [3] utilizing advanced LLM technology to remove data bias and overweighting, and [4] by using a straightforward scoring system to facilitate interpretation.

While prior research addressed questions similar to those explored in this paper, their recommendations could be limited because of their use of unsuitable metrics and by the sole use of one-dimensional Compustat® data or SIC codes to form industry clusters. Consequently, these prior studies may not have satisfactorily resolved the limitations of traditional industry classification systems. Instead, this study utilized product descriptions

from SEC 10-Q and 10-K filings to extrapolate and maximize nuanced conclusions about companies' industrial dynamics. Furthermore, this study utilized LLMs to create and quantify firm-to-firm comparisons instead of following prior literature that employed text vectorization and cosine similarity. The application of LLM technology allows company relationships to be formed on the basis of company and product descriptions instead of relying on keywords within the dataset. The result of this approach provides more accurate relationships and reduces the risk of product descriptions being improperly interpreted. Additionally, the methodology described in this paper is not static, as was the case in prior literature. Instead, the approach is highly scalable, flexible, and incorporates the addition of new data. This flexibility also allows the methodology to identify and quantify challenging, nuanced economic linkages such as customer-supplier relationships—also a subject of this study's scope. While prior literature has taken important steps towards measuring customer-supplier relationships; mainly through identifying vertical cash flows within supply chains; this study expands on those efforts by offering a more dynamic, nuanced, and reliable framework.

Table 1 of the results presents this study's summary statistics. 210 firms similar to those constituting the S&P 500 were used, ranging across various sectors, as shown in the table. Holding companies and conglomerates were intentionally included in the sample as the LLM can comprehensively weigh subsidiaries and diverse product lines within a firm's broader operational context. Including these multi-sector firms enhanced the robustness and validity of the analysis by ensuring that the methodology remained unbiased and applicable across a diverse range of business structures. The number of firms, 210, was chosen due to considerations of the expense of interacting with the OpenAI API while running the matrix. This sample was also formed from a diverse set of firms in terms of the Book-to-Market ratio (B/M), Market Value of Equity (MVE), and the Natural Logarithm of Market Value of Equity, as shown in Table 2, and a robust set of quarterly data for firms, shown in Figure 1. Tables 3 and 4 contain Ordinary Least Squares (OLS) regression models that describe the extent to which the new industry classification approach accurately measures clusters for competing and auxiliary firms, compared to SIC and NAICS codes. Table 3 answers this specifically in terms of competing firms, while Table 4 does this for firms deemed customer-suppliers by the various measures. When taking into account the weaker statistical power of models with negative LLM coefficients, Table 3 found that the AI-driven methodology was generally superior at mapping industry clusters for competing firms. This was exemplified through the positive coefficients and Z-test values of the dummy variable (the LLM measure) with their corresponding P-values of < 0.05, therefore denying the null-hypothesis of these results, in combination with the weak statistics of the models rejecting the LLM classification. The same can also be said for Table 4, but this time proving the significance of the LLM measurement at a greater scale and for customer-supplier related companies.

The paper is organized as follows. It begins with an Introductory section and a Literature Review section, together outlining the existing body of literature, this study's contributions, and its limitations. Next, in the Data section,

the data which this paper relied on is presented. The Results section explains the employed methodology for data collection and analysis, as well as presenting the specifics of the paper's findings. The paper concludes with a comprehensive summary of the key findings and discusses their broader implications. Lastly, the Conclusion section proposes directions for future research that highlights potential areas for further inquiry, as well as presenting the study's summary statistics and quantitative findings.

Literature Review

Relation to previous literature

This research aimed to create a flexible, more precise measure of industry classification and demonstrate this methodology's superior elucidation of economic links between firms in two main areas: competitors and customers-suppliers. Prior literature identified the following measures of industry: SIC codes, Compustat® segment data, internet-pulled firm data, and text-based data. Previous studies attempted to use these measures to quantify information spillover effects across industries and to find economic links and predictable returns within and across industries.

Concerning industry definitions, prior literature, including that by Guenther and Rosman, Jacobs and O'Neill, Marozzi, identified SIC codes as a relatively limited and broad measure of firm-to-firm relations (1-3). Specifically, both Guenther and Rosman (1) and Jacobs and O'Neill (2) highlighted SIC codes' lack of reliability and discrepancies across various sources: specifically, depending on the dataset (Compustat® versus CRSP for instance), SIC codes can vary. Furthermore, Marozzi (3) found that SIC codes lacked the granularity for selecting comparable firms, as they do not always group firms that are financially comparable, potentially leading to inaccuracies in financial analysis and valuation. In parallel, Ali et al., Chychyla and Kogan, and J Keil concluded that despite Compustat® segment data's robust features and decades of utilization, it presented limitations in data reliability, completeness, accessibility, and analysis capabilities (7, 8, 9). In particular, these studies point out Compustat®'s lack of data reliability compared to US Census data and SEC 10-K filings. Similarly, Du et al. highlighted that Compustat® and similar datasets commonly contained discrepancies in firm data across different sources and instead turned to SEC filings as a more reliable alternative (10).

Although Papagiannidis et al. (5) made a significant contribution to the literature by diverging from SIC codes and instead relied on internet-pulled data, their approach lacked the nuanced conclusions that text-based approaches encapsulate. In light of this, a prevailing measure researched in Hoberg and Phillips (4) utilized text-based analysis of firms' SEC 10-K product descriptions. They vectorized firm text data and applied cosine similarity to measure the similarity between companies. This methodology emerged as significant due to its ability to potentially map competitors without relying on the non-exhaustive measures of SIC codes and Compustat® segment data. Studies such as Gentzkow et al. (11) examined the text-based methodology utilized by Hoberg and Phillips (4) and concluded that similar text-based measurements of industry and firm relationship were invaluable in that they offered a more

scalable, reliable, and predictive version of traditional industry classification codes.

With respect to the predictability of returns using economically linked firms, two conclusions can be expected and have been documented: [1] no predictable returns can be made between similar companies and their announcements and [2] predictable returns can be made between groups of economically linked firms. Cohen and Frazzini concluded that stock prices do not incorporate news involving related firms (12). On the contrary, Charles et al. identified a strong predictive power for returns of similar companies, finding a long-short strategy based on the technological links of firms yielded a monthly alpha of 117-basis points (13). Despite their focus on technological links between firms, the approach outlined here of linking companies through products and services is significantly similar. Additionally, Casassus et al. (14) also showed that economic links among commodities created a source of long-term correlation between futures returns. Although this study was primarily focused on the futures and commodities markets, the evidence of return co-movements—the correlated movement of two or more entities—among linked futures highlights the inconsistency of the findings in related literature. This necessitates an addition to the literature to reach a more contemporary and accurate understanding.

Contribution to prior literature

Prior literature's use of outdated data – which does not account for the notable changes the economy has undergone in recent years - may have compromised its accuracy and applicability. In response, this analysis presents a post-Covid-19 perspective on text-based industry classification. The majority of previous studies took place roughly a decade ago with the most recent studies still being years in the past. This study offers an up-to-date analysis of the possible returns between linked companies utilizing the most precise form of measurement and most recent stock return data.

In prior literature, most studies also utilized less precise methods for measuring industry relationships. SIC codes, the former standard for identifying similar companies, are excessively broad, linking firms based solely on general industry or sector classifications. Furthermore, although NAICS codes are an advancement over SIC codes, they still present similar drawbacks. By only capturing sector or industry, these approaches overlook the myriad existence of companies competing or benefiting each other through released products and services. In contrast, this study acknowledges this and delves deeper into the comparison of competitors and linked companies, measuring them by the relation of products and services offered. This was achieved through the paper's most significant contribution: the use of large language models. Although Murdock et al. (6) presented similar methodology, by mapping the customer-supplier relationship, their study did so by identifying vertical relationships within supply chains, using cash flows.

This study utilized OpenAI's API to extract a list of all products and services listed in the examined companies' 10-Q and 10-K MD&A filings. In particular, the API is valuable as it can uniquely process prompts incrementally, analyzing input line-by-line rather than as a

single block, enabling more accurate, context-aware, and thoughtful responses compared to typical AI models. By doing this, the LLM makes even larger improvements over conventional numeric metrics such as cosine similarity, which rely strictly on textual overlap and word frequency to assess similarity. For instance, two firms might both mention the word "software" in their product descriptions - one referring to accounting software for small businesses, and the other to cybersecurity software for enterprise clients. A cosine similarity approach would register a high similarity score due to shared keywords, despite the firms operating in entirely different markets. In contrast, the LLM approach used in this study can recognize the distinct contexts of these products and assign a lower similarity score that more accurately reflects their operational separation. Once a comprehensive list of income streams is assembled, the API was utilized to compile a description of all products and services provided. In a culmination of all the compiled information, this study's LLM methodology rated companies' extent of relationship (competitors and customer-suppliers) against all other related firms in the sample. This method outputs a 210x210 similarity matrix that contains the relationship score of all firm pairings according to both types of linkages. The language model identifies customer-supplier relationships as two companies that rely on each other to buy and sell products to each other. Competitors are located by identifying companies in similar fields that have products that threaten to make another company's product either obsolete or decrease its market share. These relationship scores are used to create clusters of linked firms through a system of score thresholds.

Materials and Methods

Data

This study gathered data crucial for its investigation into the similarity between companies by first utilizing the Management's Discussion and Analysis of Financial Condition and Results of Operations (MD&A) section of SEC 10-Q and 10-K filings, spanning from Q1 2018 to Q4 2023. While the SEC sample is not as comprehensive as the return and firm statistical data utilized, it sufficiently meets the requirements for this analysis. This is because the most recent quarterly product data will contain all active product listings, meaning that there is no value in extending the SEC sample further back. In fact, expanding its start point might introduce more noise to the data sample. Product listings and descriptions were extracted from these filings, compiled, and expanded upon by the implementation of the OpenAI API to enhance clarity and robustness of product descriptions. This focused approach allows for a detailed insight into a company's financial health and product-related activities, as presented by the management. We implemented monthly logarithmic stock returns from the NASDAQ API and firms' summary statistics data from Compustat®, ranging from Q1 2015 to Q4 2024 and Q3 2014 to Q3 2024, respectively. Although this sample period contained abnormal data, caused by COVID-19's effect on the economy, these data points were intentionally included as the measurement of correlation should reliably persist as correlated firms should struggle in conjunction with each other. Additionally, even though the NASDAQ API and Compustat® sample periods did not perfectly correspond it was inconsequential, since the Compustat® data was simply used to report on the statistics of the firms in the sample while the NASDAQ API

return data was used to run the study's actual analysis. Hence, because these datasets were used for separate objectives they did not need to perfectly overlap.

Similarity Matrix

This research used a novel approach to measure industry classification by analyzing the textual content of the MD&A sections of SEC 10-Q and 10-K filings using LLM technology. The methodology focused on quantifying the relationships between firms in terms of competition as well as between their roles within each other's supply chains. To accomplish this, the product and service descriptions from the SEC filings were first extracted. Then, the OpenAI API was utilized to clean the pulled product data, fill in any missing product descriptions, and add clarity to existing descriptions in order to enable enhanced accuracy on the succeeding steps. In the next step, OpenAI's "GPT-4" LLM was instructed to compare the similarity of every firm's product descriptions, product-by-product and to quantify the strength of the following two relationships in an ordered pair: Competitors, and Customer-Suppliers. As mentioned, GPT-4 does not compute firm similarity using conventional numeric metrics such as cosine similarity, which solely rely on textual overlap to assess similarity. Instead, the model qualitatively assessed each firm's product descriptions by leveraging its advanced reasoning and semantic understanding, derived from extensive training on diverse and large-scale textual datasets. Internally, GPT-4 transforms textual information into high-dimensional numeric embeddings (underlying numerical representations capturing intricate semantic and contextual details) which it implicitly references to make holistic, contextual interpretations. While the model still internally uses text vectorization and embeddings; unlike cosine similarity; it principally relies on qualitative judgments to ultimately assign the relationships, reflecting a more nuanced textual understanding. When instructing the GPT-4 model, direct competitors were explicitly defined as companies with substantial product or service overlap competing within the same market space, thereby ensuring that widely recognized direct competitors consistently received scores at or near 5 (the maximum score). Similarly, customer-supplier relationships were explicitly described as firms that have a clear transactional relationship, in which one firm provides essential products or services directly to the other, thereby also resulting in scores approaching 5. Based on these interpretations, GPT-4 assigned each firm pair a similarity score ranging from 1 (minimal or no relationship) to 5 (strong, direct relationship). Consequently, two companies that are widely recognized as direct competitors, offering highly similar products or services within the same market space—such as Ford and General Motors—would typically receive a competitor similarity score near or equal to 5. This was confirmed by the (5,1) score these firms received in the similarity matrix. Similarly, commonly recognized customer-supplier companies—such as Nike and Foot Locker—would frequently get a customer-supplier similarity score of near or equal to 5. This assumption was corroborated by the (1,5) score in the matrix output for these firms. This qualitative scoring approach allows GPT-4 to consider nuanced product and market information beyond simple keyword matching or mathematical vector distances. Although such qualitative assessment inherently

introduces some variability compared to strictly numeric measures, this potential variability was minimized by consistently setting the model's generation "temperature" parameter, which controls the randomness and creativity of the generated text, to 0, ensuring more stable and predictable outputs. For instructions, the LLM was provided with the following guideline for the "Competitor" relationship: "Companies that have significant overlap in their products or services and compete in the same market space. For example, Coca-Cola and PepsiCo would score high as Competitors because their product portfolios are nearly identical." Next, the "Customer-Supplier" relationship was outlined as the following: "Companies where one provides products or services to the other. An example would be Apple Inc. and Foxconn, where Apple sells consumer technology supplies, and Foxconn assembles many of Apple's products." Following the guidelines for defining relationships, the LLM used its advanced transformer-based architecture to analyze the product descriptions in a nuanced manner. Unlike traditional methods, GPT-4 assesses text contextually, capturing intricate relationships without standard vectorization. It generates embeddings from deep neural layers, reflecting comprehensive language understanding. For each firm pair, GPT-4 evaluates the contextual similarity of their product descriptions and is capable of dynamically updating this analysis with new data. These evaluations were systematically arranged into a 210x210 matrix, detailing the strength of competitive and customer-supplier links, thereby providing a sophisticated and updated framework for industry classification. However, prior to running the 210x210 matrix, a sub-sample of 25 firms was created in order to run a preliminary 25x25 matrix. This sub-sample was run 13 separate times, later using these outputs to randomly select 50 matrix cells and manually validate outputs by ensuring that the assigned relationships were correct. Additionally, variances between the pairings of firms in each of the 13 outputs were compared and were sufficiently low, confirming that the LLM could generate quantified relationships with minimally varying assignments. For each firm pair, the average similarity score across all 13 runs was first computed, and then the extent to which each individual score deviated from that average was measured. This variance indicates the consistency of the LLM's similarity assignments, with lower variance reflecting more stable and reliable results. Although not part of this study's primary analysis, this served as a validation step to confirm the reliability of the LLM measurement before advancing to the full sample. This ensured the LLM was not producing erratic or substantially inconsistent results, thus avoiding potential time inefficiencies and unnecessary API usage costs. This conclusion of dependability was further backed by direct examinations of the GPT-4 LLM by Hackl et al. and Zhao et al. who have proven the reliability of the model (15, 16). In particular, the 2023 study inspected the GPT-4's ability to quantify and rate text samples in a similar text-based approach employed in this study of LLM's potential industry classification.

Specifically, the similarity assessment was set up to assign a score on a scale from 1 to 5, with intervals of 0.5, where 1 represents no strength of the given relationship and 5 indicates a high degree of that association. Scores are derived based on the contextual analysis of product descriptions, considering various factors that

denote competitive or cooperative relationships. By systematically assigning and integrating these similarity scores, the methodology provides a nuanced and dynamic map of industry relationships. This process not only enhances the precision of industry classification but also offers a deeper understanding of the economic linkages that influence market behavior and company strategy.

Clustering methodology

For the regression and industry classification metric comparison methodologies, the approach relied on the previously described similarity matrix, assignment of score thresholds, Compustat® data, and NASDAQ stock price return data. To enable historical return analysis, clusters were first created for each classification method (using the similarity thresholds from the LLM-based matrix and, for SIC and NAICS, using industry codes extracted from Compustat®). These clusters allowed firm groupings to be compared over time based on their monthly log return data. For this study's measurement, similarity score thresholds were first assigned to both relationships: 4.0 for competitors and 3.5 for customer-suppliers. These thresholds were selected after preliminary tests for each interval of 0.5 indicated they consistently balanced a sufficient cluster size and accurate identification of economically meaningful firm relationships. Specifically, a threshold of 4.0 for competitors consistently yielded clusters large enough for robust analysis, while maintaining high-confidence in relationship assignments. The slightly lower threshold of 3.5 for customer-supplier relationships was chosen to reflect the more subtle, nuanced nature of these interactions observed in the initial validation runs. To further clarify how clustering operated based on these thresholds, each pair of firms in the similarity matrix received an ordered pair score: the first number denoting the extent to which two firms are competitors, and the second number representing the strength of their customer-supplier relationship (competitor, customer-supplier). For instance, firm pairings with scores such as (4.0, 1.0), (4.5, 2.0), and (5, 2.5) would all be clustered together as competitors because each competitor score meets or surpasses the threshold of 4.0, despite their low customer-supplier scores - as that was considered irrelevant when evaluating extent of competition. However, pairs with scores such as (3.5, 2.0), (3.0, 3.0), or (2.0, 4.0) would not form competitor clusters since their competitor scores are below the threshold of 4.0, even if the customer-supplier scores are moderate to high. Conversely, firm pairs scoring (2.0, 3.5), (1.0, 3.5), and (0.5, 4.0) would be grouped into the same customer-supplier cluster, given their customer-supplier scores exceed the threshold of 3.5, regardless of their low competitor scores. A concrete example from one of this study's competitor clusters is the three-firm cluster of PepsiCo, Keurig Dr. Pepper, and Coca-Cola. Here, the PepsiCo-Keurig pairing was scored (4.5,1), the Keurig-Coca-Cola pair received a (4.5,1), and the PepsiCo-Coca-Cola pair also received (4.5,1). Thus, given that the competitor scores surpassed the 4.5 threshold, these firms formed a competitor cluster with each other. The methodology's clustering decision is reinforced by established real-world business interactions, as all three companies operate in the global beverage market and offer directly competing product lines. On the other hand, one customer-supplier cluster was the two-firm cluster of Nike and Foot Locker

where the Nike-Foot Locker pairing was scored (1,5). Given that the customer-supplier score surpassed the 3.5 threshold, these firms formed a customer-supplier cluster with each other. The plausibility of this cluster is corroborated by well-documented real-world ties: Nike is a major supplier to Foot Locker, which in turn serves as a key retail distributor for Nike products. It should also be noted that customer-supplier clusters are expected to be smaller because these relationships are typically one-to-one; while many firms engage in customer or supplier relationships, it is less common for multiple firms to be mutually linked within the same interconnected supply chain. In later iterations of this study, despite changes in the sample, this threshold system will continue to be useful even if the firm density at various score intervals drastically changes. This is because the thresholds were deliberately chosen to ensure that only clearly defined and strongly identified relationships would form clusters. Across repeated tests of the LLM's scoring, most firm pairs identified as having meaningful economic relationships consistently received scores close to or above these thresholds. The clustering of these strong relationships around scores of 3.5 to 4.0 indicates that the model is effectively separating firms with genuinely close economic connections from those with weaker or uncertain ties. In other words, because the majority of economically relevant relationships consistently score near these threshold values, the chosen thresholds reliably distinguish meaningful firm relationships from those less economically significant. Furthermore, even if the thresholds were to rise or fall by 0.5, this change's effect on the study's results would not be significantly impacted as even the slightly lower conviction relationships are distributed towards the higher end of the 0 to 5.0 range. Although lowering the threshold may slightly reduce intra-cluster correlation, the impact is likely offset by the high number of accurate relationships retained and the inclusion of some under-scored but truly correlated firms. Additionally, in the future various threshold-generated clusters for the LLM measurement will be regressed against the SIC and NAICS clusters, therefore giving the LLM measurement multiple granularities similar to the code classifications, contributing to a fairer comparison against more granular measurements.

Once thresholds were defined, firm clusters were identified accordingly. For competitor clusters, a first round of smaller clusters were identified by a passing of the score threshold between firms. Then, larger clusters were formed when each company within a cluster had a threshold-passing competitor score with at least one (in smaller clusters of two companies) or two (in larger clusters) with other companies within the same cluster. This method forms larger clusters by leveraging a transitive property where a firm competing with another is likely also a competitor to any firm that the second one competes with. While this process holds logically and broadly across industries it assumes that competitive relationships are transitive—for example, if Firm A competes with B, and B with C, then A and C may also be competitors. Hence, it can introduce some looser connections at the edge of clusters. For customer and supplier relationships, the clustering process similarly requires companies to pass a threshold score to initiate a cluster. For larger clusters, it needed strong links to at least two other firms to ensure the cluster's relevance and stability. This

prerequisite is grounded in a balance between flexibility and robustness. In smaller clusters, a single high-confidence linkage ensures meaningful firm relationships. However, as cluster size increases, the probability of false positives and chain-linking errors grows. By requiring two threshold-passing relationships in larger clusters, this methodology minimizes the risk of erroneous connections while preserving the integrity of larger clusters.

To form the SIC and NAICS code clusters, competitor clusters were first formed by seeing overlap of the industry-defined meanings of the codes' digits across the multiple granularities. For example, in the NAICS system, a 3-digit code such as 334 represents "Computer and Electronic Product Manufacturing." All firms sharing this code were grouped into a cluster, under the assumption that they compete within that specific industry segment. This process was repeated at each level (1-digit through 6-digit) to form clusters of increasing specificity as these codes are structured for firms sharing more digits to be more similar operationally. The customer-supplier clusters were formed similarly, using established industry pairings to define likely customer-supplier relationships across all granularities as these pairings historically show strong economic interdependence in common supply chains. For example, the SIC-2 specificity of "36," denoting the electronics industry, relies on "35," which represents the machinery industry.

Regression methodology

Once clusters were assigned, pulled NASDAQ monthly log returns from Q1 2015 to Q4 2024 were aggregated in a data frame for each cluster. To clarify the regression process used in the study: clusters were independently created using three different classification methods (LLM, SIC, and NAICS) where the LLM was tested against the other two methods which, in turn, served as baselines for comparison. For each resulting cluster, monthly log returns of involved firms were compiled, and pairwise return correlations among firms within each cluster were calculated and then averaged, producing a single intra-cluster correlation measure per cluster. Each cluster then became a single observation in the regression dataset, with the dependent variable as the intra-cluster correlation measure, and an independent binary variable indicating whether the cluster originated from the LLM method (coded as 1) or traditional SIC/NAICS methods (coded as 0). An Ordinary Least Squares regression then evaluated whether the LLM-derived clusters systematically exhibited stronger intra-cluster correlations than those derived from traditional classification methods. Importantly, individual similarity matrix scores (e.g., competitor or customer-supplier scores outputted by the LLM) were not directly used as regression variables. Instead, although they were integral in forming the clusters, they served no direct role in the analysis portion of this study which was simply intended to quantify the improvements or faults of the LLM. As mentioned, the LLM-based clusters were tested against SIC and NAICS benchmarks, to determine the efficacy of the LLM measurement. This comparison was done using two Ordinary Least Squared (OLS) regressions: one regression for LLM versus SIC and one for LLM versus NAICS. This regression methodology can be seen in the following governing regression equation:

$$y_{Method} = \beta_0 + \beta_1 \cdot x_1 + \epsilon \quad \text{(eq1)}$$

Within each regression, the competitor and customer-supplier clusters were independently compared using a pairwise correlation system as the basis for each comparison. This pairwise correlation system compared the returns of all pairs of firms within each cluster, calculating the degree to which their returns moved together (e.g., firm A vs. firm B, firm B vs. firm C and so on), providing a foundation for measuring the explanatory power of the LLM. Methods that use only average cluster correlations can oversimplify relationships and often lack enough observations for reliable analysis. Conversely, pairwise correlation ensures enhanced comprehensiveness and statistical power. With the basis for comparison, the OLS regression was run by implementing HC3 as a standard error estimator to account for heteroscedasticity - variations in the error terms across observations - and outputting crucial statistics to quantify the efficacy of the LLM industry measurement. While only some of these statistics were included in the regression tables due to importance, the OLS regression models supplied the following statistics for both the constant and dummy cases: model coefficient, Z-test, P-value, confidence interval, R^2, Adjusted R^2, F-statistic, and additional metrics that are less relevant in nature. Regression coefficients indicate how independent variables influence the dependent variable. In the context of this research, the coefficients represent the average of intra-cluster pairwise correlations as a means for comparison between industry classifications. The Z-statistic assesses the significance of these coefficients by measuring their deviation from the mean. Adjusted R^2 quantifies the model's fit, adjusting for the number of predictors to avoid overfitting. The P-value evaluates the probability that the observed results could occur under the null hypothesis—a default statement that there is no effect or no difference between groups or variables—with lower values suggesting stronger evidence against the null hypothesis. The F-statistic determines the overall significance of the model by comparing the variance explained by the model to the unexplained variance, confirming the collective impact of the variables.

Results and Discussion

Summary Statistics

Tables 1 and 2 and Figure 1 present the summary statistics for various financial metrics and firm characteristics for the paper's sample. In Table 1, a breakdown of firm distribution across various sectors or offices is provided. The data reveals a higher concentration of firms in the Office of Manufacturing with 52 firms, which is followed by the Office of Energy & Transportation and Office of Trade and Services with 33 and 31 firms respectively. This sectoral distribution provides insights into the primary areas of economic activity among the analyzed firms.

Table 1. Distribution of industries across the sample's firms.

Industry Sector	Number of Firms
Office of Manufacturing	52
Office of Energy & Transportation	33
Office of Trade & Services	31
Office of Technology	21
Industrial Applications and Services	21
Office of Life Sciences	19
Office of Finance	16
Office of Real Estate & Construction	9
Unknown	8
Total	210

In Table 2, the Book-to-Market ratio, Market Value of Equity (MVE), and the Natural log of Market Value of Equity for 210 unique firms over an average of 8,190 observations are shown. The discrepancy between the expected 8,400 observations (210 firms × 40 quarters) and the actual 8,190 observations is due to a combination of missing data in the Compustat® dataset, which are unavoidable, and necessary data cleaning steps. Specifically, observations with missing or invalid values in key financial fields (such as Book Value of Equity, Quarterly Shares Outstanding, or Quarterly Closing Price per Share) were excluded to ensure the accuracy of the summary statistics. Additionally, observations where Market Value of Equity or Book-to-Market ratios were zero or undefined were removed. These steps were implemented to maintain data integrity and ensure that only complete and accurate data points were included in the analysis.

Table 2. Various summary statistic metrics for the sample given in their minimum, 10th percentile, 25th percentile, 50th percentile, 90th percentile, maximum, and mean value alongside the number of observations for each metric.

	Min	10th	25th	50th	75th	90th	Max	Mean	N
Book to Market	-222.7	0.046	0.170	0.382	0.663	1.034	502.2	0.468	8189
Market Value of Equity (in millions)	0.035	90.8	588.1	3001.4	37248.1	172057.9	3522211.1	67834.2	8191
Natural log of Market Value of Equity	-3.356	4.509	6.377	8.007	10.525	12.056	15.075	8.314	8191

The data shows a wide spectrum of firm sizes and market valuations (e.g., the range of values for the Book-to-Market ratio ranges from a minimum of -222.7 to a maximum of 502.2) which reflects the diverse financial health and market valuations of these firms. The Market Value of Equity shows a broad range as well (e.g., the range of values goes from as low as 0.035 million to a staggering 3,522,211.138 million) which highlights the vast differences in the market capitalizations of firms in the dataset. The mean Market Value of Equity stands at 67834.2 million, portraying a positively skewed distribution as it is significantly driven upwards by larger firms. This skewness is further evidenced by the mean MVE of 67834.2 million that is significantly higher than the 0.035 million minimum MVE value and the 3001.4 median MVE value. Furthermore, the positive skewness is highlighted by the Natural log of Market Value of Equity, with a mean of 8.314 and a median of 8.007.

As seen in Figure 1, the data consistency across firms with varying quarterly reporting periods. A substantial portion of the firms, 168 to be precise, have 41 quarters of data, while 12 firms have 40 quarters and 6 firms have 39. An insignificant number of firms fall below 39 quarters.

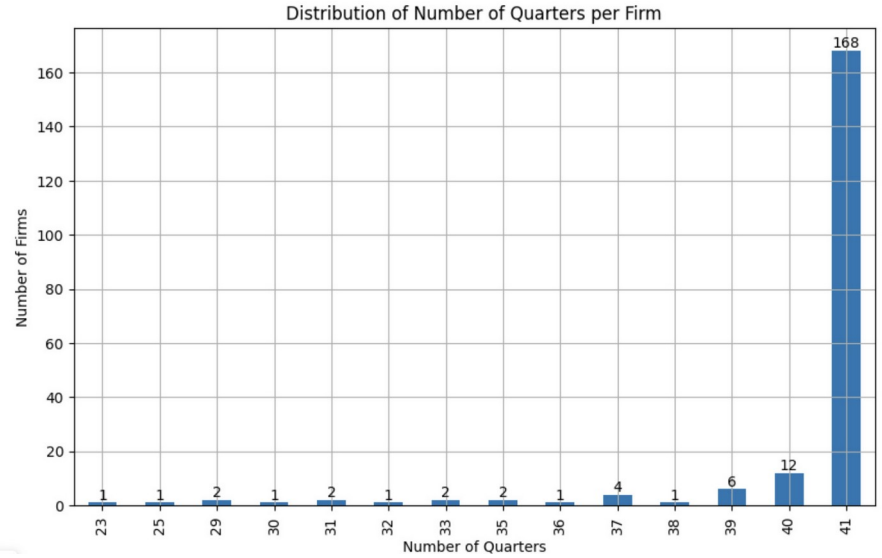

Figure 1. Distribution of the number of quarters per firm for the sample's firms.

The firms containing a significantly lower number of quarters can be explained by the cleaning process that was performed on the utilized sample. In this case, repeated product descriptions were removed, causing firms with small product portfolios to be populated with data across a small number of quarters. This distribution suggests a common trend in the data reporting frequency among the majority of the firms, which could be indicative of a standardized financial reporting cycle and which aids in a more uniform analysis across the dataset.

Regression Results

The regression analysis in Table 3 and 4 assessed the efficacy using a Large Language Model (LLM)-based classification system compared against various granularities of traditional SIC and NAICS classifications in identifying economic relationships between firms. The 1-digit to 6-digit nomenclature referenced in Tables 3 and 4 corresponds to the level of granularity in the NAICS and SIC classification systems. While this system is inherent to the NAICS and SIC code structures, the terminology of "x-digit" when referring to the granularity of the code is used in related literature such as Hoberg and Phillips (4). These classification systems follow a hierarchical structure where each additional digit refines the specificity of the industry definition. At the 1-digit level, industries are broadly categorized (e.g., manufacturing or finance). As the granularity increases, each subsequent digit provides a more precise definition, narrowing the focus to specific subsectors, industry groups, and detailed business activities. For instance, a 1-digit NAICS code might identify the manufacturing sector, while a 4-digit code could specify computer and electronic product manufacturing, and a 6-digit code could pinpoint semiconductor manufacturing. In this study, these varying granularities allow for a comprehensive evaluation of how well the LLM-based classification system performs in capturing firm relationships compared to traditional NAICS and SIC measures at different levels of detail. By comparing results across these levels, the extent to which the LLM's classification improves intra-cluster correlation for both competitor and customer-supplier relationships is assessed. However, it should be noted that the LLM measure is evaluated using a single threshold for each relationship whereas the industry code-classifications had multiple levels of granularity, inherently placing the LLM at a disadvantage in the regression analysis. Despite this limitation, the LLM still demonstrated superior performance, suggesting that future iterations with multiple score thresholds could yield even greater accuracy and robustness. The results from these tables show the higher predictive power of the LLM approach. Table 3 focuses on competitor relationships. It showed overall improvement in identifying competitive dynamics when using the LLM approach. This is illustrated by the LLM's coefficients for NAICS and SIC classifications at different granularity shown in Table 3.

For instance, looking at the 1-digit NAICS and SIC level, the LLM method achieves a 0.204 coefficient for the NAICS model and 0.198 coefficient for the SIC model, demonstrating this measurement's 20.4% and 19.8% respective improvement in intra-cluster correlation compared to baseline. Furthermore, these results correspond to statistically significant P-values, adjusted R^2 values of 0.228 and 0.168, and Z-statistics of 59.04 and 37.514, respectively. This rejects the null hypothesis, and confirms the LLM's statistically significant superiority at these granularities and proves that the LLM is a good fit for identifying competitor relationships. At the 2-digit NAICS and SIC levels, the LLM exhibited notably positive coefficients (0.194 and 0.097, respectively) with P-values of 0, rejecting the null-hypothesis of these models and suggesting the ability to discern subtle competitive nuances missed by traditional models. The LLMs have Z-statistics of 27.153 and 7.013. It should be noted that even though

the 2-digit NAICS maintained a good model fit, with a 0.132 adjusted R^2, the SIC 2-digit model was a relatively poor fit with an adjusted R^2 of 0.022. The NAICS-3-digit model with a coefficient of 0.073, shown in Table 3, is the last model that fits LLM-superiority pattern with a 4.886 Z-statistic, a near 0 P-value, and 0.011 adjusted R^2. While this demonstrates a significant improvement over the NAICS measurement, it is likely a poor model fit due to the adjusted R^2 value.

Table 3. OLS regression outputs for the competitor clusters comparison for the LLM measurement against both the SIC and NAICS measurements.

Method	Coefficient (constant)	Coefficient (x_1)	z (x_1)	P>|z| (x_1)	Adjusted R^2	F-statistic	N
NAICS-1-digit	0.039	0.204	59.040	0	0.228	3486	11804
NAICS-2-digit	0.096	0.194	27.123	0	0.132	735.6	4852
NAICS-3-digit	0.232	0.073	4.886	0	0.011	23.87	1998
NAICS-4-digit	0.387	-0.050	-2.298	0	0.004	5.28	1198
NAICS-5-digit	0.432	-0.104	-4.411	0	0.017	19.45	1074
NAICS-6-digit	0.598	-0.228	-8.266	0	0.080	68.33	776
SIC-1-digit	0.067	0.198	37.514	0	0.168	1407	6972
SIC-2-digit	0.211	0.097	7.025	0	0.022	49.35	2202
SIC-3-digit	0.404	-0.068	-3.013	0.003	0.007	9.08	1148
SIC-4-digit	0.570	-0.185	-6.792	0	0.053	46.13	814

The LLM method exhibited a decline in efficacy at granularities beyond the NAICS 3-digit and SIC 2-digit models for competitor classification. For the NAICS 4, 5, and 6-digit models, respective negative LLM coefficients of -0.05, -0.104, and -0.228 were found, displaying the LLM's inferior ability to cluster competitive firms at these granularities. While Z-statistics and P-value still suggested statistical significance for results at these granularities, the adjusted R^2, F-statistic, and number of observations were weaker than those of models with positive LLM coefficients. For instance, the NAICS 4 and 5-digit models showed particularly weak adjusted R^2 values of 0.004 and 0.017, respectively, although the 6-digit granularity held an acceptable value of 0.080. Furthermore, the 4 and 5-digit NAICS models had low F-statistics of 5.28 and 19.45, respectively, while the NAICS 6-digit model remained significant with a F-statistic of 68.33.

Evaluating the SIC 3 and 4-digit models for competitor classification, even though the LLM method had significant negative coefficients of -0.0683 and -0.1846 with strong Z-statistics of -3.013 and -6.792, respectively, the other statistical tests were not strong. These granularities demonstrate lower than preferred adjusted R^2 values and number of observations. The SIC 3-digit model even contained 1 non-rejected null-hypothesis through a 0.003 P-

value and a poor F-statistic of 9.08. The 3 and 4-digit SIC models showed 0.007 and 0.053 for adjusted R^2 with considerably lower observation counts at 1,148 and 814, accordingly. Although an adjusted R^2 of 0.053 is not ideal, it can suffice. However, when compared to the observation counts of 1 and 2-digit SIC granularities (which had a minimum of ~ 2,000 observations each up to a maximum of 11,804 observations for the NAICS-1-digit model) the more granular models displayed much less statistical power. Overall, the regression results in Table 3 demonstrate the LLM method's superiority in identifying competitor relationships across the less granular half of all competitor models (i.e, particularly 1 and 2-digit levels for both NAICS and SIC). The results show the significantly superior statistical power of the LLM method across the models at these granular levels and thereby validates the use of LLM classification over NAICS and SIC code alternatives. Table 4 focuses on customer-supplier relationships, showing that the LLM measure displays positive, significant coefficients up until the last digit of granularity for both the NAICS and SIC clusters.

Table 4. OLS regression outputs for the customer-supplier clusters comparison for the LLM measurement against both the SIC and NAICS measurements.

Method	Coefficient (constant)	Coefficient (x_1)	z (x_1)	P>|z| (x_1)	Adjusted R^2	F-statistic	N
NAICS-1-digit	0	0.231	412.893	0	0.497	170,500	172,382
NAICS-2-digit	0.001	0.236	128.661	0	0.488	16,550	17,390
NAICS-3-digit	0.002	0.181	80.777	0	0.394	6525	10,028
NAICS-4-digit	0.256	0.153	1.941	0.052	0.036	3.767	78
NAICS-5-digit	0.049	0.096	5.077	0	0.058	25.77	406
NAICS-6-digit	1	-0.595	-16.028	0	0.937	256.9	20
SIC-1-digit	0.001	0.2423	147.479	0	0.506	21,750	21216
SIC-2-digit	0.004	0.2261	63.029	0	0.417	3973	5560
SIC-3-digit	0.032	0.1369	10.282	0	0.146	105.7	618
SIC-4-digit	0.909	-0.53	-5.07	0	0.565	25.71	22

In particular, the NAICS granularity measures ranging from l-digit to 5-digits showed coefficients of 0.231, 0.236, 0.181, 0.153, and 0.096, respectively, which is in tandem with near-zero P-values and respectable adjusted R^2, F-statistics, and observation counts. This excludes the NAICS 4-digit model because it displayed unusually weak statistical power, failing to reject the null-hypothesis with a P-value of 0.053, an adjusted R^2 of 0.036, a F-statistic of 3.767, and a comparatively inadequate 78 observations. The overall strength of the LLM statistics for NAICS granularities of 1, 2, 3, and 5-digits shows its robustness in capturing inter-firm customer-supplier linkages. Even when the flawed

NAICS 4-digit model is included, these 5 models yield an average adjusted R^2 value of 0.295, indicating a model fit and affirming its reliability in predicting economic relationships. Similarly, the 1, 2 and 3-digit SIC models in Table 2 yield positive coefficients of 0.242, 0.226, and 0.137, respectively, with near-zero P-values and an average adjusted R^2 value of 0.356 across the three models. Furthermore, the 1, 2, and 3-digit SIC models hold robust F-statistics, ranging from 105.7 in the 3-digit model to 21,750 in the 1-digit model. The three SIC models also show strong Z-statistics, ranging from 10.282 in the SIC 3-digit model to 147.479 in the 1-digit model.

In general, these findings collectively affirmed the LLM's advanced capability over SIC and NAICS codes at the majority of granularities for identifying customer-supplier relationships. It should be noted that for NAICS 6-digit and SIC 4-digit granularities, the LLM returned inferior intra-cluster pairwise correlations with coefficients of -0.595 and -0.53, accordingly. Although the NAICS 6-digit and SIC 4-digit models rejected the null hypothesis while holding statistically significant Z-statistics, adjusted R^2 values, and F-statistics, they exhibited minimal observation counts of 20 for the NAICS 6-digit model and 22 for the SIC 4-digit model. The low observation counts decrease their statistical power and diminish the usefulness of the highly granular versions of each method. This is because industry classifications' value must also take into account the number of firms that they are able to cluster with accuracy, which the NAICS-6-digit and SIC-4-digit models are unable to do. Consequently, in future iterations of this study, a quantity-adjusted OLS regression will be implemented so that the number of firms within a cluster is measured as part of its statistical success.

Overall, superior performance of the LLM method across a meaningful number of granularities for both competitor and customer-supplier relationships, provides compelling evidence of its ability to more accurately and dynamically map industry classifications. This is particularly true for customer-supplier classification. This superior performance is reflected across various statistical measures in Table 2, which provides conviction that the LLM approach offers a significant improvement for economic linkage analysis in academic research and practical applications. Furthermore, at the more granular (and therefore, seemingly more accurate) NAICS and SIC variations, the observation count drops off steeply, which shows the insufficient robustness – less statistical power - of those highly granular measurements that outperformed the LLM measurement in terms of intra-cluster correlation to cluster a robust set of firms.

Strategic and Financial Applications

A critical contribution of the LLM-based methodology lies in its enhancement of industry mapping and firm classification. Traditional classification systems such as NAICS and SIC codes provide static and often outdated industry boundaries that may be unable to capture dynamic market conditions, technological innovations, or shifts in business strategy. The proposed LLM-based system, by leveraging qualitative, context-aware textual analyses of firm product descriptions, provides dynamic industry maps that can update in near-real-time as firms evolve or emerge. Practically, users can generate an updated

similarity matrix for the entire S&P 500, identifying competitor and customer-supplier clusters based on the provided thresholds (e.g., ≥ 4.0 for competitors). Analysts could then regularly review these clusters to identify emerging industries, monitor structural industry shifts, and accurately group firms for comparative research and investment analysis.

Beyond improved industry mapping, the methodology offers significant value for portfolio diversification and risk management. Conventional wisdom often views competition in terms of zero-sum relationships. However, the LLM-based classification recognizes more nuanced competitive relationships. For example, consider Coca-Cola and Pepsi: investors might traditionally allocate similar portfolio weights independently, leading to unintended sector concentration. In practice, users can apply this method by reviewing the similarity matrix, identifying highly related firms (score ≥ 4.0), and adjusting portfolio weights to avoid over-concentration. Specifically, investors could reduce their allocations (e.g., from 6% each to 3% each), effectively treating closely correlated firms as a single combined exposure. Such nuanced adjustments help manage portfolio risks and sector concentration effectively, even when both firms independently benefit from market growth. This approach can similarly apply to subtler relationships identified uniquely by the LLM.

The applications of this method extend further into direct risk assessment. Financial analysts and risk managers could leverage these relationship scores by first identifying firm clusters based on high similarity scores. Subsequently, they can analyze historical co-movements within these clusters, incorporate identified clusters into scenario analyses, and perform stress tests considering correlated shocks. This explicit recognition of nuanced firm linkages yields more accurate and economically meaningful risk assessments.

Additionally, the LLM-based firm classification significantly benefits strategic merger and acquisition targeting. Practitioners can use the similarity matrix to identify natural acquisition targets or partners with highly synergistic scores. Firms with strong product or operational overlaps identified by high competitor or customer-supplier scores could then be systematically assessed for potential strategic alignment, informing targeted M&A decisions. Similarly, supply chain managers can use this data to identify potential suppliers or customers not evident from standard industry classifications, strengthening supply chain robustness.

Finally, from a broader economic research perspective, the LLM-based classification framework provides researchers with a powerful tool for analyzing sector spillover effects, firm co-movement, and market transmission effects with greater precision. Researchers can systematically select firm clusters based on the similarity scores, validate these clusters through economic analysis of historical returns, and subsequently study industry-wide impacts or inter-firm ripple effects more accurately than with traditional methods. This facilitates more nuanced insights into market behaviors, significantly contributing to both academic literature and practical market understanding.

Limitations

Despite this paper's addition to prior literature, the approach still presented limitations, including the current consistency and accuracy of large language models, data anomalies, product disclosure inconsistencies, and lack of consideration for revenue stream percentages. Although LLMs have been taking the world by storm, this technology's novelty presents issues around consistency and researchers' ability to depend on its use without errors. Furthermore, given the continuous evolution of LLMs, replicating this study could prove challenging, as it would be difficult to exactly match the deployed LLM. Although it was minimized in this study through a set temperature of 0, because the GPT-4 LLM's outputs are generative, parameters such as temperature can introduce variation in assigned scores, meaning repeated runs may yield slightly different similarity judgments for the same firm pairs. Additionally, much of the utilized dataset coincides with an altered economy due to COVID-19 and its implications. These data points were intentionally included as this measurement of correlation should reliably persist as correlated firms should struggle in conjunction with each other. However, this abnormal financial period still presented the possibility for skewed results. In a similar context of dataset limitations, not every company in the sample has NAICS codes, and there were minor gaps in the return data. These deficiencies in the NAICS code and NASDAQ return data, while unavoidable, still impacted the analysis.

An unpreventable limitation of this study is the disclosure rationale behind firms' product descriptions. For starters, it is safe to assume that many defense contractors and companies linked to government activity are less likely to disclose all new product information as it could have implications for a nation's military efforts or private affairs. Although this only affects the aggregation of product descriptions needed to run the similarity matrix, it is impossible to account for or eliminate. Similarly unavoidable, because the LLM, SIC, and NAICS classification systems differ in structure and interpretability, each required a slightly different clustering approach. While efforts were made to standardize the methodology as much as possible, these underlying differences in how each system defines firm relationships inherently limited the use of a fully uniform clustering process.

The final considerable limitation of this analysis is the lack of consideration for revenue percentage in the similarity matrix. In an attempt to negate this, LLM technology was implemented to add further detail to product descriptions, adding some clarity to how significant any product could be in the broader context of firms' revenue. Furthermore, the LLM methodology takes into account the context of each product within a company's product portfolio, partially removing this issue. However, this AI-driven contextualization is not perfect and cannot fully replace the insights that would come from having detailed revenue data by product. In particular, this study draws connections between the similarity of products and the type of business companies conduct. For example, this is problematic in the case of one product being a trivial percent of one business' revenue but meanwhile being a significant portion of another company's revenue. This is troublesome because these companies would not be truly competing because of the differing importance of the products to each firm's bottom line. However,

to the similarity matrix, these products and companies would possibly draw a similarity signal no different than if both companies' products shared equal weight in terms of revenue percentage. As mentioned, this is partially counteracted by the proposed methodology's contextualization of products' places in a firm's broader business, but nevertheless, this does not completely avoid this limitation.

Conclusion

This research aimed to develop a more refined and accurate measure of industry classification by using LLM technology to analyze textual data from company product descriptions in SEC filings. The study sought to determine: [1] how this novel methodology compares to traditional classification systems such as SIC and NAICS in identifying and quantifying relationships between competitors and customer-suppliers; and [2] how this impacts investor strategies and market understanding.

The study diverged from previous literature in its methodology, data, and future prospects. In contrast to prior papers' use of cosine similarity and one-dimensional metrics, this investigation is unique in implementing LLMs as the comparison methodology. While SIC codes, NAICS, and Compustat® segment data have long been used for industry classification, they face challenges such as static structures and discrepancies across data sources, as highlighted in prior literature. To mitigate this, the study's method provided a more recent and novel analysis to understand the competitive and auxiliary landscapes within the market. This approach captured a more detailed view of industry relationships by directly analyzing firm-level product descriptions from SEC filings without the concerns of tokenizing textual-data. Shahmirzadi et al. comment on this issue, showing that text vectorization faces challenges such as high dimensionality, lack of semantic understanding, inefficiency in incremental updates, and reliance on subjective tuning (17). However, neural models, such as the GPT-4 LLM utilized in this study, excel at capturing contextual semantics and outperform simpler methods in in-domain tasks and short text analysis. Because of this, the conclusions not only enhance current industry classification methods but also present immense future prospects by setting the stage for the inevitable broader AI integration within the literature.

Data for this study's initial analysis were gathered from the SEC MD&A 10-Q and 10-K filings which were then combined with NASDAQ logarithmic stock price return data for regression calculation and Compustat® firm data for summary statistics calculations. This blend of company-reported data and objective return and statistics data provides a robust basis for the analysis. The regression results, which were formed by comparing various industry classifications' intra-cluster correlation using monthly logarithmic returns, indicated that the explanatory power of clusters with an LLM origin versus SIC or NAICS codes was greater on intra-cluster correlation for a majority of granularities.

The statistical backbone of this research consisted of OLS regression models. These models were applied to test the impact of LLM - measured clusters versus SIC and NAICS using intra-cluster pairwise correlations to measure the success of the various measures. Most notably, the majority of results in support of the LLM methodology had P - values < 0.05

with corresponding positive z-test values, thereby rejecting the null-hypothesis and demonstrating the significance of the findings. On the other hand, many of the results that refuted this industry classification either exhibited poor model fits or generally weak statistical power. Because of this, the LLM methodology was identified as generally better at clustering firms and producing high intra-cluster correlations than both SIC and NAICS codes across the various granularities and firm relationships, especially for the customer-supplier linkage. The specific superiority of the LLM for the customer-supplier models is crucial as it shows distinct improvement for more nuanced firm relations than the simple competitor linkage. Demonstrating this was one of this study's main objectives. Also, it is essential to recognize that these results provide a foundational basis, and as the reliability and capabilities of LLMs advance, the outcomes are expected to enhance correspondingly.

This research makes several contributions to the existing body of literature and market practices while opening numerous avenues for future research. By introducing a more precise, AI-driven method of industry classification based on product descriptions from SEC filings, this research challenged and extended traditional models such as SIC and NAICS codes that have shown limitations in capturing the nuanced relationships between firms such as the customer-supplier link explored in this paper. Beyond classification, this methodology enables dynamic industry mapping, more precise portfolio diversification, and improved risk assessment by identifying economically linked firms. Practitioners can use the methodology and its outputs to adjust allocations, detect correlated firm clusters, and uncover strategic M&A or supply chain opportunities, demonstrating the model's practical value across finance and strategy.

The study presents significant promise for the future of the literature. AI is undeniably the next major expansion of academia and general quantitative and qualitative measurements. Since this study has shown that AI is capable of successfully quantifying complex similarities and linkages, these findings pave the way for future studies. Future studies can modify the proposed methodology (via enhanced or expanded procedures and improved LLM iterations) such that they eventually achieve a highly accurate industry classification, all while utilizing the same base technology explored in this study. While future iterations of this study will incorporate changes such as using the S&P 500 or an equivalently robust sample and testing new variables, future researchers could expand on this methodology by applying it to a larger dataset or investing more nuanced relationships beyond competitors and customer-suppliers dynamics. Researchers could also refine the methodology by narrowing the scope of product relationships by incorporating the percentage of firms' revenue that is contributed by specific product portfolios. This would allow for more nuanced relationship scoring. Moreover, this approach could be refined to include machine learning algorithms that incorporate new market data and additional public firms in order to predict shifts in industry dynamics and to provide novel arbitrage opportunities, as well as to update portfolio diversification. Lastly, future research should consider longitudinal studies to track the evolution of industry classifications over time and their power to predict firm performance and economic resilience. Such

studies would significantly enrich the models used by investors and policymakers understanding of industry dynamics and alike.
enhance the predictive accuracy of financial

Abbreviations

SIC: Standard Industry Classification, NAICS: North American Industry Classification System, SEC: Securities and Exchange Commission, AI: Artificial Intelligence, LLM: Large Language Model, API: Application Programming Interface, MVE: Market Value of Equity, B/M: Book-to-Market ratio, Min: Minimum, Max: Maximum, CRSP: Center for Research in Security Prices, MD&A: Management Discussion and Analysis, NASDAQ: National Association of Security Dealers Automatic Quotation system, GPT: Generative Pretraining Transformer, HC-x: Heteroscedasticity consistent

Notations

y_{Method} - The dependent variable of the pairwise intra-cluster correlation within each cluster
β_0 - The constant term of the intra-cluster correlation for the baseline group (SIC or NAICS cluster depending on the model and $x_1=0$)
β_1 - The difference in intra-cluster correlation between LLM clusters and baseline clusters
x_1 - The dummy variable and it indicates whether a given cluster is from the LLM method ($x_1=1$) or a baseline method ($x_0=0$)
ϵ - The error term and captures the variability in the dependent variable, random noise, or unexplained factors affecting intra-cluster correlation
N - number of observations
z (x_1) - The Z-statistic for the LLM measurement in a given model
P>|z| (x_1) - the P-value of the Z-statistic for the LLM measurement in a given model

Supplementary file

210 x 210 similarity matrix

References

1. Guenther, D. A., Rosman, A. J. (1994). Differences between Compustat® and CRSP SIC codes and related effects on research. *Journal of Accounting and Economics*, 18(1), 115-128. https://doi.org/10.1016/0165-4101(94)90021-3

2. Murdock, M., Ngo, T., Richie, N. (2022). Beyond SIC Codes: Mergers of Related Firms. *Quarterly Journal of Finance and Accounting*, 60(3/4 (Summer & Fall 2022)), 1-44. https://www.jstor.org/stable/27224938

3. Hackl, V., Müller, A. E., Granitzer, M., Sailer, M. (2023). Is gpt-4 a reliable rater? Evaluating consistency in gpt-4's text ratings. *Frontiers in Education*, 8. https://doi.org/10.3389/feduc.2023.1272229

4. Lee, C. M., Sun, S. T., Wang, R., Zhang, R. (2019). Technological links and predictable returns. *Journal of Financial Economics*, 132(3), 76-96. https://doi.org/10.1016/j.jfineco.2018.11.008

5. Jacobs, G., O'Neill, C. (2003). On the reliability (or otherwise) of SIC codes. *European Business Review*, 15(3), 164-169. https://doi.org/10.1108/09555340310474668

6. Keil, J. (2017). The trouble with approximating industry concentration from Compustat®. *Journal of Corporate Finance*, 45, 467-479. https://doi.org/10.1016/j.jcorpfin.2017.05.019

7. Ali, A., Klasa, S., Yeung, E. (2008). The limitations of industry concentration measures constructed with Compustat® data: Implications for finance research. *Review of Financial Studies*, 22(10), 3839-3871. https://doi.org/10.1093/rfs/hhn103

8. Papagiannidis, S., See-To, E. W., Assimakopoulos, D. G., Yang, Y. (2018). Identifying industrial clusters with a novel big-data methodology: Are SIC codes (not) fit for purpose in the internet age? *Computers & Operations Research*, 98, 355-366. https://doi.org/10.1016/j.cor.2017.06.010

9. Cohen, L., Frazzini, A. (2008). Economic links and predictable returns. *The Journal of Finance*, 63(4), 1977-2011. https://doi.org/10.1111/j.1540-6261.2008.01379.x

10. Casassus, J., Liu, P., Tang, K. (2012). Economic linkages, relative scarcity, and commodity futures returns. *Review of Financial Studies*, 26(5), 1324-1362. https://doi.org/10.1093/rfs/hhs127

11. Hoberg, G., Phillips, G. (2016). Text-Based network industries and endogenous product differentiation. *Journal of Political Economy*, 124(5), 1423-1465. https://doi.org/10.1086/688176

12. Marozzi, M. (2013). A method to address the effectiveness of the SIC code for selecting comparable firms. *Electronic Journal of Applied Statistical Analysis*, 6(2). https://doi.org/10.1285/I20705948V6N2P186

13. Chychyla, R., Kogan, A. (2014). Using XBRL to conduct a large-scale study of discrepancies between the accounting numbers in Compustat® and SEC 10-K filings. *Journal of Information Systems*, 29(1), 37-72. https://doi.org/10.2308/isys-50922

14. Gentzkow, M., Kelly, B., Taddy, M. (2019). Text as Data. *Journal of Economic Literature*, 57(3), 535-574. https://www.aeaweb.org/articles?id=10.1257/jel.20181020

15. Du, K., Huddart, S., Jiang, X. D. (2023). Lost in standardization: Effects of financial statement database discrepancies on inference. *Journal of Accounting and Economics*, 76(1), 101573. https://doi.org/10.1016/j.jacceco.2022.101573

16. Zhao, M., Li, F., Cai, F., Chen, H., Li, Z. (2024). Can we trust llms to help us? An examination of the potential use of gpt-4 in generating quality literature reviews. *Nankai Business Review International*. 16(1), 128-142. https://doi.org/10.1108/nbri-12-2023-0115

17. Shahmirzadi, O., Lugowski, A., Younge, K. (2019). Text Similarity in Vector Space Models: A Comparative Study. *18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, Boca Raton, FL, USA. 659-666. https://doi.org/10.48550/ARXIV.1810.00664