Next-Generation Global Financial Business Networks via Multimodal Embedding Models and Look-ahead Bias Control [*]

Peng Yifeng
yifengpeng@link.cuhk.edu.cn

June 30, 2025

Abstract

Accurately charting the fast-evolving web of inter-firm ties is crucial for forecasting risk transmission, informing regulation and uncovering investment opportunities, yet conventional industry codes or return-based co-movements remain too coarse, slow and often biased to meet this need. To overcome these shortcomings, this project unifies state-of-the-art LLMs embeddings with contrastive multimodal fusion of text, images, audio and structured fundamentals, and embeds explicit time-aware masking and benchmark tests to eliminate look-ahead bias while objectively selecting the best representation model for finance. The resulting framework yields high-fidelity, time-stamped vectors for a globally comprehensive firm universe, enabling the construction of a dynamic business network whose edge strengths transparently reflect economically meaningful relationships. By delivering a scalable pipeline, a rigorously validated multimodal embedding space and a bias-controlled network dataset, the study is expected to advance theoretical benchmarks for financial representation learning and provide practitioners and regulators with more accurate tools for risk monitoring, strategic analysis and alpha generation.

[*] This article's English writing and grammar have been polished with the assistance of AI tools such as ChatGPT and Google Gemini Pro.

Contents

1 Introduction

1.1 Research Background

Accurately characterizing the intricate and dynamic networks of inter-firm relationships within the global economy is paramount for understanding the evolution of market structures, predicting pathways of financial risk contagion, and formulating effective investment and regulatory strategies. Prevailing methodologies for constructing business networks, however, exhibit significant limitations.

Traditional approaches reliant on static industry classification standards (e.g., Global Industry Classification Standard (GICS), Standard Industrial Classification (SIC), North America Industry Classification System (NAICS)) inherently possess high-level aggregation, low granularity, and update latencies. These shortcomings render them inadequate for capturing the increasing heterogeneity, dynamic changes, and complex interplay of cross-sector collaborations and competitions prevalent in modern economic systems [20]. Such classification methods frequently mask substantial intra-industry variances between firms concerning their actual business emphases, positioning within supply chains, technological development trajectories, and target customer segments. Consequently, they fail to unveil the unique and true economic risk exposures and potential opportunities these firms face. Moreover, early attempts to infer economic linkages using financial market data, such as stock return correlations, are susceptible to distortions from systemic market noise and non-fundamental factors. This susceptibility can lead to the identification of spurious co-movements, thereby limiting the reliability of these methods in precisely delineating substantive economic connections [10, 29].

The ascent of deep learning-based LLMs and their resultant embedding representations, presents a compelling new paradigm for surmounting previously identified limitations [7]. Through in-depth analysis of extensive unstructured textual information publicly disclosed by corporations-such as annual reports, investor conference call transcripts, and press releases-these technologies facilitate the extraction of profound semantic associations that extend beyond traditional classification schemes and market-derived data [13, 22]. Embedding models, in essence, project discrete textual objects (e.g., corporate business descriptions) into high-dimensional continuous vector spaces, wherein semantically cognate enterprises are positioned in close proximity. This methodological capacity to capture intricate semantic relationships, synonymy, and context-dependent economic linkages-often indiscernible to conventional keyword matching or bag-of-words models-establishes a robust foundation for constructing more economically meaningful global business networks [5].

Despite the notable advancements in text-based embedding methodologies, an exclusive reliance on a singular textual modality may result in the oversight of pivotal information that a firm transmits to the market. Contemporary corporate communication and information disclosure increasingly exhibit multimodal characteristics; non-textual data, encompassing images (such as product photography, factory visuals, and satellite imagery), video (including promotional films, technological demonstrations, and press conferences), and audio (for instance, executive interviews and vocal content from conference calls), frequently harbor unique business signals that textual data alone cannot comprehensively articulate or capture [8, 47].

Though progressing in the use of embedding technologies for the construction of global financial business networks, several key challenges persist, directly shaping the core inquiries of this research. A primary concern arises from the prevalent reliance on earlier-generation embedding models [5]. These models may exhibit inherent limitations in capturing contemporary business dynamics and nuanced semantic distinctions, and the cut-off dates of the training corpus can introduce issues of data timeliness. Compounding this is the critical, yet often underestimated, problem of **look-ahead bias**. Models pre-trained on corpus that include future information may inadvertently leverage this ”foreknowledge” when analyzing historical data, leading to distorted representations of past networks [18, 27, 39].

The development of robust, systematic strategies to counteract this bias, particularly within the specific context of financial network construction, remains notably underdeveloped. Furthermore, a discernible gap exists in the rigorous, finance-application-oriented comparative assessment of current leading-edge embedding models. This includes prominent examples such as OpenAI’s text-embedding-3 series, Nvidia’s NV-Embed-v2, and Voyage AI’s Voyage-3-Large, for the task of building domain-specific business networks. The optimal methodologies for effectively fusing multimodal data to maximize informational synergy, while concurrently designing adaptive fusion mechanisms that accommodate the heterogeneous quality and relevance of diverse data modalities, also remain unresolved research questions [32].

Collectively, these unaddressed critical issues underscore an urgent imperative: the development of a next-generation framework for global financial business network construction. Such a framework must be characterized by enhanced precision and robustness, the sophisticated integration of multimodal information, and the incorporation of effective mechanisms for bias control.

1.2 Research Motivation

While initial advancements in the the construction of financial business networks through NLP and AIGC are evident, prevailing methodologies demonstrate marked deficiencies in depth, scope, accuracy, and robustness. These shortcomings demonstrate such approaches fall short of the sophisticated analytical demands posed by contemporary, intricate global economic systems. Such deficiencies impede a comprehensive understanding of the deep-seated economic linkages between firms and substantially restrict the application potential of these approaches in pivotal financial decision support. Hence, the central motivation for this research is the systematic resolution of following interconnected and critical research gaps.

1.2.1 Limitations of Information Acquisition and Biases

Significant deficiencies pervade the data foundations upon which current network constructions are based. A critical limitation stems from the prevalent reliance on datasets with restricted scope, such as those confined to large, publicly traded corporations or pre-processed textual data from specific repositories. This practice frequently results in the omission of Small and Medium-sized Enterprises (SMEs) and economies within emerging markets-entities integral to the global economic fabric-thereby yielding networks that lack comprehensive global representation. Moreover, even when LLMs are employed to generate corporate descriptions, a conspicuous absence of systematized validation and governance protocols persists [5].

The efficacious and precise distillation of standardized, information-dense corporate characterizations from voluminous, heterogeneous, and unstructured primary textual sources-including, for instance, nuanced details within annual report footnotes, complex contractual disclosures, and the qualitative narratives of Management Discussion & Analysis sections-remains a formidable and largely unresolved impediment. These foundational inadequacies at the data stratum directly and substantially impinge upon the ultimate quality and dependability of any subsequent network construction endeavors.

1.2.2 Insufficient Adaptability of Model Evaluation

The rapid proliferation of embedding model technologies underscores a critical gap: current research lacks systematic, comparative evaluations of the latest generation for the sophisticated task of constructing financial business networks [32].

Current researches often resorts to deploying general-purpose models or specific advanced models without the framework of a rigorous, multi-dimensional, domain-specific evaluation benchmark designed for financial applications. This benchmark is indispensable and should encompass assessments of economic plausibility, network topological stability, and performance in relevant downstream financial tasks, including risk contagion analysis and return forecasting. This deficiency in standardized, finance-centric evaluation frameworks fosters an element of arbitrariness in model selection, thereby undermining confidence in the financial interpretability and practical utility of the resulting networks.

1.2.3 Systematic Disregard of Multimodal Signals

Prevailing researches are characterized by an overwhelming reliance on textual data, thereby systematically overlooking the critical economic signals embedded within the increasingly rich multimodal data landscape of corporate disclosures (e.g., product launch videos, factory operation imagery, executive interview audio, patent diagrams) [6,8,51]. Such alternative information frequently furnishes complementary, or even counter-textual, insights that are either elusive to or inadequately articulated by textual modalities alone; for instance, image recognition can reveal a company's daily operational status, while voice emotion analysis can offer insights into executive confidence. A significant lacuna in extant research is the absence of an innovative, multimodal analytical framework capable of synergistically fusing these heterogeneous data sources and projecting them into a unified semantic space to cohesively represent enterprises and their interrelations. This oversight-the failure to integrate multimodal information, which inevitably yields a partial depiction of inter-firm economic linkages.

1.2.4 Reliability Challenge of Look-ahead Bias

The utilization of pre-trained large models, inclusive of embedding models, for the analysis of historical data introduces a pervasive and methodologically critical challenge: look-ahead bias [18,27,39]. During their pre-training phase, these models are exposed to information post-dating the temporal points they are subsequently tasked to analyze or predict. This exposure can result in the construction of historical networks imbued with anachronistic linkages, severely distorting historical actualities and thereby undermining the validity of empirical analyses predicated thereon, such as the backtesting of trading strategies or the assessment of historical event impacts [24,41].

While extant research acknowledges this pervasive issue, and initial mitigation techniques like information masking have been explored [18], the development of systematic, verifiable bias control strategies that effectively balance the imperative of bias elimination with the critical need for pertinent information preservation remains a significant lacuna. Simplistic interventions, such as the wholesale masking of all data perceived as ”future” information, risk an excessive forfeiture of valid and valuable signals. Consequently, a core research imperative lies in the meticulous design of sophisticated, context-aware, and granular information masking or rectification mechanisms, specifically tailored to the nuanced characteristics of financial texts. Equally critical is the establishment of robust frameworks for the quantitative evaluation of these mechanisms' efficacy, a step paramount to ensuring the demonstrable reliability of research conclusions.

1.3 Research Objectives

Predicated on the pressing unresolved critical issues outlined above, this research endeavors to construct and rigorously validate an innovative framework. This framework, leveraging state-of-the-art AI technologies-encompassing advanced LLMs, multimodal learning paradigms, and sophisticated embedding models-is conceived for the comprehensive construction and analysis of global financial business networks. The efficacy of this framework is contingent upon a synergistic application of systematic data processing, stringent model evaluation protocols, pioneering multimodal fusion techniques, and robust bias control mechanisms. The overarching ambition is to achieve an unprecedented standard of precision, comprehensiveness, dynamism, and reliability in delineating inter-firm economic linkages on a global scale. To realize this, the research delineates five closely interrelated objectives.

- **Establishment of a systematic and scalable high-quality global database of corporate descriptions**

  This involves developing an automated pipeline that integrates advanced LLMs, sophisticated prompt engineering, Retrieval Augmented Generation (RAG) techniques, and rigorous data validation processes. This pipeline is designed to extract standardized, temporally consistent, and information-rich corporate descriptions from diverse global textual sources-including annual reports, conference call transcripts, news articles, and regulatory filings-thereby significantly broadening the coverage to include SMEs and firms within emerging markets.

- **Establishment of a finance-domain-adaptive evaluation benchmark for embedding models**

  This objective entails the design and implementation of a multi-dimensional evaluation framework. Such framework will incorporate criteria such as economic plausibility, network topological stability, and performance on downstream financial tasks, to systematically compare the efficacy of leading-edge embedding models within the specific context of financial network construction. The careful outcome of this stage will be the identification of the most suitable model, or a synergistic combination of models, for subsequent network construction.

- **Development of an innovative multimodal information fusion framework**

  A joint embedding architecture, potentially leveraging contrastive learning or cross-modal attention mechanisms would be developed during research. The architecture aims to effectively project heterogeneous information sources-encompassing text, images, video, audio-derived sentiment, and structured financial data-into a unified semantic space. An integral part of this objective is the exploration of adaptive weighting mechanisms to optimize fusion efficacy, ultimately leading to the generation of more holistic and nuanced representations of inter-corporate relationships.

- **Design and Validation of Effective Look-Ahead Bias Control Strategies**

  This involves the meticulous development of fine-grained Named Entity Recognition (NER) and context-aware masking techniques, customized for the financial domain. A systematic quantitative assessment of the inherent trade-offs in various strategies-balancing the imperative of eliminating look-ahead bias against the critical need for preserving valid economic signals-will be undertaken. This phase is expected to culminate in the proposal of robust best-practice methodologies for bias mitigation in financial AI applications.

- **Empirical Validation of the Practical Economic Value of the Enhanced Network**

  The high-fidelity, multimodal global business network developed will be leveraged for in-depth investigations into several critical areas:

  - Exploring the lead-lag effects in global stock returns as driven by multimodal signals
  - Enhancing the early prediction of corporate M&A by leveraging network structure and multimodal features
  - Elucidating the propagation mechanisms and pathways of ESG risks and opportunities within the constructed network
  - Conducting a network-based assessment and stress-testing of global supply chain resilience

The successful realization of these objectives is anticipated to not only substantially advance the research frontiers within the FinTech domain but also to furnish robust, innovative, and practically valuable methodologies and analytical instruments.

1.4 Research Significance/Contribution

This research endeavors to construct a global financial business network predicated upon the latest embedding models and the fusion of multimodal data. The anticipated outcomes are poised to engender substantive impacts spanning theoretical innovation, practical applications, and broader societal value.

1.4.1 Theoretical Contributions

- **Establishing a Benchmark Framework for Embedding Models in Finance**

  A critical gap in current research is the absence of a systematic, multidimensional comparative assessment-encompassing network topological stability, economic linkage interpretability, and downstream task predictive performance-for the proliferating array of novel embedding models specifically within the context of financial business network construction. This research would try to addresses this deficiency by proposing a rigorous, task-oriented evaluation benchmark and selection rubric for the financial research community. This framework would transcends generic NLP metrics, thereby establishing a foundational methodology for subsequent, related inquiries.

- **Pioneering Innovative Methodologies for Multimodal Financial Data Fusion**

  Current financial network research exhibits a pronounced dependence on unimodal textual data. This study introduces and will empirically validate a novel Joint Embedding architecture-potentially exploring techniques such as contrastive learning or shared-specific space decomposition-designed to effectively map heterogeneous information sources, including text, images, video, vocal sentiment, and structured financial data, into a unified semantic space. The methodological innovation resides in its explicit modeling of the complementarity and redundancy across diverse modalities. Through a bespoke adaptive fusion mechanism, this approach is poised to substantially enhance the comprehensiveness and accuracy of business network representations, thereby charting new theoretical avenues for financial information processing.

- **Providing Systematic Bias Mitigation Strategies for Time-Series Financial Research**

  Look-ahead bias presents a formidable challenge when employing pre-trained large models for historical financial analysis. Addressing this, the present research designs and quantitatively evaluates bias control strategies predicated on NER and context-aware masking, specifically tailored to the idiosyncrasies of financial texts. Beyond enhancing the reliability and veracity of the historical networks constructed herein, this investigation critically contributes an operational and evaluable methodological framework for the responsible application of large pre-trained AI models in financial time-series analysis. This framework holds considerable import for other financial AI research reliant on historical data modeling.

1.4.2 Practical Contributions

- **Empowering Investment Decisions: Uncovering Alpha and Managing Novel Risks**

  The network facilitates more precise identification of lead-lag effects across firms and industries, allowing for the earlier capture of momentum propagation, thereby informing the optimization of pair trading and sector rotation strategies. It supports the development of novel alternative data factors derived from multimodal network centrality and dynamic linkage-strength metrics, potentially constituting new sources of alpha for quantitative investment strategies. Furthermore, event-driven strategies are substantially improved; inter-firm linkages identified by the network enhance the capability to predict M&A likelihoods, discern potential targets and beneficiaries, and assess the contagion scope of events such as ESG controversies, thereby enabling more efficacious risk management and opportunity capitalization.

- **Enhancing Corporate Strategic Insight and Supply Chain Resilience**

  For corporate entities, the research offers a panoramic and dynamic view of the competitive landscape. This is achieved through real-time monitoring of global competitors-including unlisted firms and entities in emerging markets-encompassing their product launches, strategic market adjustments, and evolving customer relationships, thereby cultivating enhanced competitive intelligence. The capability for precise mapping of global supply chains is also a significant contribution. This allows for the assessment of latent risks in supplier and customer portfolios (e.g., financial distress, ESG vulnerabilities, geopolitical exposures) by integrating multimodal information. Consequently, a quantitative analysis of supply chain fragilities and the identification of alternative sourcing or diversification opportunities become feasible.

- **Assisting Regulators in Systemic Risk Monitoring and Assessment** The developed methodology and network data also furnish regulatory bodies with powerful analytical tools. The quantification of financial contagion risk is enhanced, as the multi-layered inter-firm network facilitates more accurate simulations of risk propagation pathways and intensity under various shocks, such as interest rate fluctuations, industry-specific scandals, or geopolitical conflicts. Identification of systemically important nodes-firms or industry clusters pivotal to global financial or supply chain stability-is enabled by the multimodal network topology, supporting more focused regulatory scrutiny and robust stress-testing. The framework allows for the assessment of macroeconomic policy impacts, specifically by analyzing the potential structural reshaping of global business networks attributable to particular industrial policies or trade agreements.

1.4.3 Social Value and Long-term Impact

- **Facilitating Quantitative ESG Factor Assessment and Sustainable Development**

  The network analysis framework developed in this research is poised to significantly advance the quantitative assessment of ESG factors and promote sustainable development. A core capability of this framework lies in the meticulous quantification of ESG spillover effects; by tracing and measuring the propagation of ESG risks, such as environmental pollution incidents, and opportunities, like the adoption of green technologies, through intricate business networks, it furnishes critical empirical evidence concerning the tangible impact of ESG-focused investments. Furthermore, the framework offers a robust mechanism for verifying corporate ESG commitments. The integration of multimodal data facilitates a rigorous cross-validation of publicly disclosed ESG information against observable corporate behaviors within the network-including supply chain partner selections and associations with adverse news-thereby promising a substantial enhancement in market transparency. Ultimately, by fostering a more precise and nuanced understanding of the transmission mechanisms for ESG risks and opportunities, this research is anticipated to catalyze a more effective allocation of capital towards enterprises genuinely committed to sustainable practices, thereby directly supporting the United Nations Sustainable Development Goals.

- **Advancing Responsible AI Application and Development in Finance**

  This research makes a substantive contribution to the advancement of responsible AI application and development within the financial domain. The direct confrontation and systematic resolution of look-ahead bias-a critical challenge in deploying large models for financial analysis-coupled with a dedicated exploration of interpretability in multimodal fusion, inherently operationalizes key principles of responsible AI. A significant outcome is the establishment of a methodological exemplar for the deployment of sophisticated AI models in high-risk, time-sensitive financial decision-making contexts. This is achieved by pioneering best practices concerning transparency, reliability, and rigorous bias control, offering a valuable blueprint for the field. Moreover, the research lays a foundational framework for long-term inquiry. The developed multimodal network construction and analysis architecture possesses inherent iterative value and scalability. As data sources continue to burgeon and models undergo rapid evolution, this framework is positioned to serve as a durable infrastructure and a novel analytical paradigm for future FinTech investigations, which, in the long term, are expected to contribute significantly to enhancing information efficiency and optimizing resource allocation within financial markets.

2 Literature Review

2.1 Business Network Construction

The accurate delineation of intricate and dynamic economic interconnections among enterprises, realized through the construction of Business Networks (BNs), is of paramount importance for comprehending market structures, information dissemination pathways, risk contagion phenomena, and the formulation of efficacious financial strategies [1, 14]. However, the precise and comprehensive capture of these linkages, particularly within the context of an increasingly globalized economy, perennially constitutes a core challenge within financial economics research. The extant body of literature has predominantly investigated business network construction through three principal avenues, each of which is encumbered by specific limitations, thereby underscoring the imperative for the innovative methodological approaches advanced in this study.

2.1.1 Traditional Industry Classifications Approaches

The most straightforward methodologies for constructing BNs have relied upon standard industry classification systems, such as the SIC, the GICS, or the NAICS. These frameworks operate on the foundational assumption that enterprises categorized within the same industry segment share substantive economic linkages, a premise that has led to their widespread adoption owing to their inherent simplicity and the ready availability of requisite data [15]. Nevertheless, this core postulation of intra-industry homogeneity encounters escalating challenges in the contemporary economic landscape. Such static and often overly broad classification schemes prove inadequate in capturing the progressively pronounced heterogeneity within industries, an issue highlighted by Hoberg & Phillips [20]. The high degree of product line specialization, nuanced differentiation in market positioning, and considerable diversity in supply chain architectures mean that firms nominally situated within the same traditional industry can, in reality, confront markedly disparate competitive environments and risk exposures [5]. Illustratively, an enterprise like Tesla, despite its classification within the automotive sector, exhibits profound divergences from traditional internal combustion engine manufacturers such as Ford, particularly concerning technological trajectories, target clientele, and supply chain dependencies. Traditional classifications, however, frequently aggregate such distinct entities, thereby severely constraining the precision and explanatory capacity of networks derived from these methods.

2.1.2 Market Data-Based Approaches

To surmount the inadequacies of static classifications, attention of researchers has shifted towards leveraging market-derived data, notably the correlation or co-movement of stock returns, as a means to discern economic linkages [17]. The underlying premise posits that firms subject to comparable risk exposures or possessing substantive economic ties should manifest a greater degree of synchronicity in their stock price fluctuations [3]. Nonetheless, this approach is beset by two principal challenges:

- **Signal-to-Noise Ratio Issues**

  Equity markets are inherently susceptible to a multitude of non-fundamental influences, such as pervasive market sentiment, liquidity shocks, and macroeconomic noise. These factors can precipitate spurious co-movements among firms that lack genuine economic connections, thereby introducing substantial noise and diminishing the reliability of the resultant networks [4].

- **Ambiguity and Omission in Relationship Typing**

  Correlations in stock returns often fail to elucidate the specific nature of the economic ties (e.g., competitive, supply-chain, or shared risk exposure). Furthermore, their capacity to identify linkages characterized by less efficient market pricing or time-lagged information transmission-such as intricate cross-industry or cross-regional supply chain connections is demonstrably limited [10, 29].

With the advancement of NLP technologies, constructing BNs using firms' publicly disclosed textual information releases has emerged as a promising approach.

- **Early Textual Analysis: Bag-of-Words (BOW) Models**

  Pioneering work by Hoberg and Phillips introduced the Textual Network Industry Classification (TNIC) method. By analyzing the "Business Description" section of U.S. public companies' 10-K filings using BOW models, they dynamically identified competitive relationships among firms [19–21]. This approach offers finer granularity and more dynamic industry structures than traditional classifications. However, BOW models inherently ignore word order and context, struggling with synonyms (e.g., "vehicle" vs. "automobile") and polysemy (e.g., "security" meaning cybersecurity, production safety, or financial securities)-a significant limitation in information-dense financial texts. Moreover, TNIC heavily relies on the U.S.-specific 10-K format, limiting its applicability in international markets lacking standardized disclosure formats and languages. Firms' strategic discretion in disclosure, including selective omission or ambiguity, further complicates text-based relationship identification [13].

- **Modern Embedding Models**

  To overcome BOW limitations, research has shifted toward deep learning-based embedding models, especially Transformer-based architectures like BERT and GPT series. These models map text into high-dimensional vector spaces, capturing complex semantic and contextual relationships. Breitung and Müller notably advanced this field by leveraging GPT-3 to generate standardized historical business descriptions for over 60k global firms, addressing cross-national disclosure heterogeneity [5]. Using OpenAI's text-embedding-3-small and large models, they computed semantic similarities between firm descriptions to construct the first large-scale, time-varying global business network based on embeddings. Their findings demonstrate that embedding-based networks outperform traditional BOW methods in capturing industry and country homogeneity, relationship overlap, and in downstream financial tasks such as predicting global stock return lead-lag effects and merger targets.

  Nonetheless, despite these breakthroughs, challenges remain:

  - **Data Coverage and Quality Limitations**

    Deficiencies in data coverage and quality persist. B&M acknowledge shortcomings in their approach concerning the representation of small-capitalization stocks, particularly in the earlier segments of their sample period, and in certain Asian and African markets-an issue partly attributable to the difficulties of the GPT-3 tokenizer with specific languages such as Chinese.

- **Look-ahead Bias Risk and Mitigation Efficacy**

  The risk of look-ahead bias and the efficacy of current mitigation strategies remain critical concerns. B&M explicitly address the potential for LLM-based embedding models, whose training data extended to September 2021, to introduce such bias. For instance, a model might inadvertently leverage knowledge of a future merger event present in its training data, thereby artificially inflating embedding similarity between the involved firms prior to the actual event. Drawing upon methodologies akin to those in Glasserman & Lin and Kim et al., B&M attempted to mitigate this by masking firm and product names using spacy's transformer-based NER model [18, 25]. Their findings indicate that while masking had a marginal impact on lead-lag effect applications, it significantly attenuated the predictive power for M&A deals. This dichotomy suggests that look-ahead bias is indeed a salient issue and that masking offers a degree of remediation, yet it simultaneously risks expunging genuine economic signals.

- **Neglect of Multimodal Information Value**

  The value of multimodal information remains largely untapped in this domain. Existing research, including the foundational work by B&M and broader surveys on LLMs in finance, has predominantly centered on extracting value from textual data. Yet, corporate information disclosure is increasingly characterized by its multimodal nature. Non-textual data, such as charts and tables within annual reports, product imagery, videos from press conferences, and audio from executive earnings calls, often convey critical information that text alone may not fully capture or articulate-for instance, managerial sentiment, nuanced product innovations, or the visual emphasis of strategic priorities.

2.1.3 Summarize of BNs Reviews

In summation, despite considerable advancements in methodologies for constructing BNs, extant research-inclusive of the most recent approaches leveraging LLMs and embedding models-continues to exhibit notable research lacunae that directly inform the design of this study.

Deficiencies in data breadth and depth persist, with current networks demonstrating inadequacies in global market coverage (particularly concerning emerging markets and SMEs) and in the processing of multilingual, heterogeneous data. The systematic assessment and efficacious control of look-ahead bias, a critical challenge introduced by large pre-trained models, remain underdeveloped. Furthermore, a lag exists in the model evaluation and selection processes; there is a distinct absence of systematic benchmarking and comparative analysis for the latest generation of embedding models specifically within the task of financial network construction. The predominant reliance on unimodal textual analysis in current research signifies a failure to effectively integrate the rich informational value inherent in increasingly prevalent multimodal data sources. Finally, the capacity for automated and precise differentiation of various economic relationship types (e.g., competition, supply, customer) within business networks is still in its nascent stages.

This research endeavors to address these delineated challenges through the introduction of cutting-edge LLM and embedding technologies, the development of an innovative multimodal fusion framework, the design of superior look-ahead bias mitigation strategies, and the execution of systematic model evaluations and relationship-type differentiations. The ultimate aim is to construct a global financial business network characterized by enhanced precision, comprehensiveness, dynamism, and reliability.

2.2 Embedding Model in Finance

To address the limitations of traditional keyword-based or statistical methods in capturing the complex semantic relationships within financial texts, researchers have increasingly turned to deep learning-based embedding models. These models aim to map discrete textual units (words, sentences, and even documents) into high-dimensional, continuous vector spaces, positioning semantically similar texts in close proximity within this space [30, 31]. This representation provides a robust analytical foundation for understanding unstructured financial data such as company business descriptions, news sentiment, and regulatory filings, serving as a critical technological underpinning for the construction of more precise and dynamic financial business networks.

The evolution of embedding technologies has been significant. Early static word embedding models (Word2Vec/GloVe) were capable of learning semantic relationships between words (e.g., *king* - *man* + *woman* *queen*). However, they generated a unique vector for each word, failing to account for polysemy (e.g., the multiple meanings of ”bank”), a significant drawback in the term-rich and often ambiguous financial domain.

The introduction of the Transformer architecture revolutionized the field of NLP, leading to context-aware embedding models such as BERT and the GPT series [7, 12, 43]. By pre-training on large-scale corpora, these models can generate dynamic vector representations based on the specific context of a word, substantially improving the understanding of deep semantic meaning within text. For instance, the model can differentiate between ”bear” in ”bear market” and the animal ”bear.” To better compare longer text segments such as company business descriptions, Transformer-based sentence embedding models were developed. These models generate fixed-dimension vector representations for entire sentences or paragraphs, directly measuring semantic similarity between texts and avoiding the information loss associated with simple word vector averaging [35].

Recognizing the differences between general-purpose corpora and the linguistic characteristics of the financial domain, researchers have further developed domain-specific embedding models. FinBERT, for instance, fine-tunes BERT on financial news and company reports, significantly improving performance on tasks such as financial sentiment analysis and named entity recognition [2]. Similarly, BloombergGPT utilizes the vast amount of high-quality financial data accumulated by Bloomberg to provide enhanced language understanding and generation capabilities tailored to financial markets [48]. These domain adaptation efforts aim to enable embedding models to better capture the precise meanings of financial terms and industry-specific contextual information.

The expansion of model scale and training datasets has recently ushered in a new generation of highly capable embedding models. Prominent among these are API-accessible services [26, 45]. These models generally exhibit enhanced zero-shot or few-shot learning capabilities, superior multilingual support crucial for constructing global business networks, and an improved capacity for processing longer text sequences. Concurrently, potent open-source alternatives, including large T5 variants and the latest multilingual Sentence Transformer models, offer the research community valuable, customizable options that bolster transparency and reproducibility [33, 36].

By calculating vector similarities between corporate business descriptions or other relevant texts (e.g., product introductions, news reports), embedding models can identify potential economic linkages-such as competitive relationships, supply chain partnerships, or technological/thematic similarities-with greater granularity and dynamism than traditional methods based on industry codes or keyword co-occurrence [5, 10, 20]. Networks constructed from such embedding similarities have demonstrated considerable economic value in applications like predicting M&A activities, explaining stock return co-movements, and analyzing information spillover effects [28].

Despite this immense potential, applying the latest embedding models to construct global financial business networks, particularly across different historical periods, confronts formidable challenges. These challenges directly motivate the core objectives of this research:

- **Look-ahead Bias**

  A fundamental methodological hurdle in using pre-trained models for historical analysis is look-ahead bias. Models pre-trained on data extending beyond a specific historical point may inadvertently incorporate future information, such as subsequent events, terminological evolution, or corporate developments-when representing past company states. This can lead to the construction of historical networks with anachronistic connections or overestimated relationship strengths. While masking entity-specific information like company names offers a partial mitigation, such straightforward approaches may not fully prevent models from leveraging their learned, broad contextual knowledge (potentially indirectly linked to future firm performance) to, in effect, ”travel through time”. Systematically detecting, quantifying, and effectively mitigating this bias to ensure the authenticity of historical networks represents a critical technical challenge (corresponding to research object 4 of this study).

- **Lack of Standardized Evaluation and Selection**

  A deficit in standardized evaluation and selection protocols further complicates model choice for the specific task of financial business network construction. Current research often relies on generic NLP benchmarks or limited downstream task evaluations, lacking a comprehensive assessment framework tailored to financial network building. Such a framework should encompass the economic plausibility of network structures, interpretability of economic relationships, and performance across multiple relevant downstream tasks. The performance of different models can vary substantially in capturing diverse types of economic linkages or processing texts from different languages and industries. Systematically comparing the latest models and establishing robust evaluation criteria is therefore a key anticipated contribution of this research (corresponding to research object 2).

- **Interpretability and the ”Black Box” Problem**

  The ”black box” nature of embedding models poses another significant hurdle. While the high-dimensional vectors generated by these models effectively encode semantic information, their internal operational mechanisms often lack transparency. In financial applications, knowing that two companies are proximate in vector space is insufficient; understanding the reasons for this proximity-whether due to product similarity, shared customer bases, or common risk exposures-is critical for risk management, investment decisions, and regulatory oversight. Current embedding methods offer limited capabilities in furnishing such interpretability [37, 38].

- **Computational Cost and Scalability**

  Computational cost and scalability present practical barriers. Constructing dynamic global business networks necessitates processing textual data for vast numbers of firm-pairs (potentially millions or more) across multiple time points. Applying large embedding models to computations of this magnitude demands substantial resources, posing challenges for both research and real-world deployment and compelling a trade-off between model performance and efficiency [25].

In essence, while embedding models offer revolutionary opportunities for constructing financial business networks, they also highlight critical research gaps concerning look-ahead bias control, standardized model evaluation, interpretability, and the insufficient integration of multimodal information. Against this backdrop, this study aims to propel the field forward by introducing state-of-the-art models, developing systematic evaluation and bias mitigation methodologies, and exploring the potential of multimodal fusion. An in-depth examination of these challenges naturally leads to a discussion of the potential and hurdles of multimodal analysis in finance.

2.3 Multimodal Analysis in Finance

The evolution of information technology and the diversification of data sources have increasingly highlighted the insufficiency of analytical paradigms reliant solely on textual or structured data for comprehensively capturing complex market dynamics and corporate behaviors within the financial domain. Corporations increasingly disseminate critical signals not only through financial statements and textual disclosures but also via a variety of other media, including images, videos, and audio, such as executive presentations [8]. Multimodal analysis involves the integration and examination of heterogeneous data from disparate modes-textual, visual, auditory, and numerical-has emerged as a significant frontier in financial information processing and decision support. This approach seeks to transcend the limitations inherent in unimodal information, aiming to elicit deeper and more holistic insights [34, 49]. The multimodal capabilities of LLMs are particularly noteworthy, extending their application beyond text to other data forms . In finance, this is advantageous for integrating varied data sources such as news articles, financial statement data, and market charts .

A core challenge in multimodal learning resides in the effective fusion of heterogeneous data and the extraction of meaningful joint representations. Researchers have explored various fusion strategies, including early fusion (at the data layer), late fusion (at the decision layer), and, more prominently, intermediate fusion (at the feature layer) [16]. Among different strategies, joint embedding techniques are particularly pivotal, endeavoring to map data from different modalities-such as textual descriptions from corporate reports, product imagery, financial charts, and audio features from executive speeches-into a shared semantic space [11]. This enables information from diverse origins to be compared, correlated, and analyzed within a unified framework. Recent researcher introduce transformer-based cross-modal attention mechanisms into financial analysis, allowing models to dynamically focus on the interactions between different modalities of information. For example, in the joint analysis of financial report text and tabular data for fraud detection, such models can learn the associations between textual descriptions and anomalous financial indicators [42, 44].

In specific financial applications, multimodal analysis has already demonstrated its distinct value. In investment analysis, for instance, attempts have been made to combine textual reports, financial indicators, and visual charts to construct more comprehensive corporate profiles, with the objective of enhancing the accuracy of stock return predictions [40]. Within the realm of risk assessment, studies have shown that integrating textual and financial data with non-verbal cues extracted from management presentation videos-such as facial expressions and vocal tone variations-can more effectively predict credit default risk or identify signals of financial distress [9, 50]. Recent advancements also encompass the utilization of multimodal LLMs for the direct comprehension and analysis of complex financial documents containing text, images, and charts [23, 46]. These models are capable of identifying crucial chart-based information within financial reports and even generating preliminary analytical opinions, showcasing potent cross-modal understanding capabilities.

Despite the advancements achieved by multimodal analysis in various financial tasks, its application to the specific objective of constructing large-scale, dynamic global financial business networks remains nascent and is confronted by distinct challenges and inherent limitations.

- **Limited Focus on Inter-firm Relationships**

  A primary impediment is the predominant research orientation within multimodal finance towards assessing individual company characteristics, intrinsic value, or risk profiles, rather than systematically leveraging multimodal signals to identify and quantify the economic linkages between firms. For instance, while multimodal information is utilized to evaluate standalone corporate entities, its potential to delineate crucial interconnections-based on signals such as visual similarity in products, shared branding elements across corporate imagery, or mentions of competitors and partners in audio-visual executive communications-is largely underexplored [40]. The central, and as yet inadequately addressed, challenge lies in architecting multimodal representation learning methodologies explicitly engineered to capture the nuanced spectrum of inter-enterprise relationships, including similarity, competition, and complementarity.

- **Data Acquisition and Alignment Challenges**

  While textual data, such as annual reports, benefit from a relative degree of standardization, the procurement of corporate images, videos, and high-fidelity audio data on a global scale-particularly historical datasets-is an endeavor fraught with difficulty. These challenges arise from disparate formats, pervasive inconsistencies, and intrinsic linguistic and cultural variances. Critically, the precise temporal and entity-level alignment of these unstructured multimodal datasets with corresponding corporate entities and associated textual information represents a formidable technical prerequisite for ensuring the accuracy of any resultant network construction. This alignment process, in itself, constitutes a non-trivial undertaking of considerable complexity.

- **Effectiveness and Interpretability of Fusion Mechanisms**

  The effectiveness and interpretability of fusion mechanisms designed to integrate these diverse data streams present open research questions. The development of robust algorithms, such as sophisticated joint embedding techniques or advanced cross-modal attention mechanisms, capable of effectively amalgamating heterogeneous financial data-which often comprises sparse visual or auditory signals alongside dense textual or numerical information-while simultaneously amplifying relationship-indicating signals, remains an area requiring significant investigation. Beyond efficacy, the inherent complexity of multimodal models, typically surpassing that of their unimodal counterparts, often culminates in a ”black box” characteristic. This opacity makes it exceedingly challenging to elucidate precisely how a specific modality (e.g., a product image or an executive’s audio segment) influences the formation of a particular link within the network.

Consequently, while multimodal analysis undoubtedly offers a richer and more nuanced lens through which to understand individual enterprises, the systematic and scalable utilization of multi-source information-encompassing textual, visual, and auditory data to construct and validate more precise and comprehensive global financial business networks remains a formidable and largely uncharted research territory. The multimodal fusion framework proposed in this study is specifically conceived to address these critical challenges.

2.4 Literature Review Conclusion

This review has systematically charted the current landscape, advancements, and inherent constraints in constructing global BNs, understanding global commercial interrelations, applying embedding models, and integrating multimodal data. The review above reveals that despite notable progress, several critical research lacunae demand attention. Traditional approaches to network construction, reliant on industry classifications or stock return correlations, frequently prove too coarse and static to adequately capture the complex, dynamic, and transnational economic ties defining contemporary economic systems. While text-analytic methods, particularly those leveraging embedding models, represent a significant step forward, they are not without their own set of persistent challenges. These include limitations in data coverage, especially concerning global markets, Small and Medium-sized Enterprises, and multilingual contexts. Moreover, the vexing issue of look-ahead bias, introduced by pre-trained models, remains difficult to entirely eradicate. There is also a discernible lack of systematic evaluation frameworks for the newest generation of embedding models, and a persistent difficulty in accurately distinguishing between diverse types of economic relationships, such as competitive versus cooperative links. Compounding these issues, current research paradigms predominantly overlook the rich informational content embedded in increasingly prevalent multimodal data sources, such as images and audio-visual materials, thereby failing to fully exploit multimodal fusion analysis for enhanced network representation. Finally, much of the existing scholarship on global business relations remains confined to specific geographical regions or offers only static perspectives, struggling to depict the dynamic evolution of global networks and their intricate mechanisms in phenomena like risk contagion and information diffusion.

Hence, the research proposed herein is directly predicated on these identified gaps, aiming to surmount the limitations of existing studies through a series of methodological innovations. This proposal is committed to developing enhanced data acquisition and processing pipelines. By harnessing the latest LLMs in conjunction with sophisticated prompt engineering techniques, the goal is to efficiently extract standardized, high-quality corporate descriptions with comprehensive global coverage. A crucial component involves the systematic evaluation and comparison of cutting-edge embedding models tailored to the task of financial business network construction, alongside the establishment of a robust quantitative assessment framework. Furthermore, this study will design and implement an innovative multimodal fusion analytics framework. Through joint embedding techniques, this framework will integrate textual, visual, video, and vocal data, thereby capturing a richer spectrum of inter-firm relational signals. Addressing the critical issue of temporal validity, more effective strategies for mitigating look-ahead bias-such as advanced Named Entity Recognition and context-sensitive masking protocols-will be proposed and validated to ensure the authenticity of historical network constructions. The practical economic value of the enhanced global business network thus developed will be comprehensively ascertained through empirical validation across several key financial application scenarios. These include, but are not limited to, global stock return predictability, M&A event forecasting, the analysis of ESG spillover effects, and the assessment of supply chain resilience. By confronting these challenges and achieving these research objectives, this study anticipates delivering a significantly improved methodological toolkit for understanding complex global economic interdependencies, thereby offering substantial contributions to both academic scholarship and financial practice.

3 Methodology Design

3.1 Data Acquisition and Preprocessing

The empirical foundation of this research proposal will be constructed from multi-source, multimodal, enterprise-level data spanning major global financial markets. The temporal scope is preliminarily established from 2000 to 2024, a period deemed sufficient to support robust dynamic network analysis. Data acquisition will rigorously adhere to principles of comprehensiveness, reliability, and accessibility.

3.1.1 Data Sources

- **Core Textual Data Sources**

  Regulatory filings represent an indispensable cornerstone for standardized, audited information.

  - **United States:** Annual reports (10-K), quarterly reports (10-Q), and other pivotal disclosures (e.g., 8-K, S-1) for publicly traded companies will be directly sourced from the U.S. Securities and Exchange Commission's (SEC) EDGAR database[1]. Particular emphasis will be placed on sections rich in business descriptions and strategic insights, notably Item 1 (Business), Item 1A (Risk Factors), and Item 7 (Management's Discussion and Analysis - MD&A). The SEC's Financial Statement and Notes Data Sets will also serve as a valuable supplementary source for structured textual and numerical information.
  - **Global Markets:** For other principal global markets, statutory disclosure documents in local languages will be acquired via interfaces from commercial data providers (e.g., LSEG[2] for international reports) or, where feasible, directly from national securities regulatory bodies' websites.

  Investor relations materials offer further crucial textual data sources.

  - **Earnings Call Transcripts:** Transcripts of publicly listed companies' earnings conference calls, replete with detailed management discussions and Q&A on performance, strategy, and outlook, will be obtained from specialized providers such as FactSet[3], Refinitiv Eikon, Bloomberg, or S&P Capital IQ.
  - **Company Annual Reports (Non-Regulatory Versions):** PDF versions of annual reports, often more reader-friendly and graphically rich than regulatory filings, will be sourced from corporate investor relations websites.
  - **Press Releases & Corporate News:** These will be acquired through APIs of commercial data vendors or by direct web scraping of corporate websites and major financial news outlets (e.g., MarketWatch, Reuters), adhering to ethical data collection practices.

  Standardized historical business descriptions from providers like S&P Capital IQ, FactSet, Bloomberg, and Refinitiv will be obtained for supplementation and cross-validation,

[1] https://www.sec.gov/search-filings
[2] https://www.lseg.com/en/data-analytics
[3] https://www.factset.com/

proving particularly useful when addressing historical data gaps or specific non-English regional data.

- **Structured and Auxiliary Data Sources**
  - **Standardized historical financial statement:** Balance sheets, income statements, cash flow statements will be acquired from established databases.
  - **Market data:** Stock prices, trading volumes, and market capitalization, will be sourced from the aforementioned providers or specialized vendors like CRSP[4].
  - **Analyst data:** Encompassing earnings forecasts and ratings, will be obtained from sources like I/B/E/S via Refinitiv.
  - **ESG data:** ESG Scores, sub-component indicators, and controversy information, will be sourced from mature data providers such as Sustainalytics, MSCI, Refinitiv ESG, or Bloomberg.
  - **Fundamental company information:** Registration details, addresses, industry classifications, and key executive data, will be retrieved from primary databases and potentially cross-verified using resources like OpenCorporates.
- **Multimodal Data Sources** The incorporation of multimodal data is intended to capture signals potentially missed by text-only analysis.
  - **Visual Data:** Key images will be extracted from company annual reports, official websites, and product brochures. These include product photographs, corporate logos, organizational charts, market share diagrams, and images of factories or facilities. Exploratory avenues include the potential leveraging of publicly available or commercial satellite imagery (e.g., via AWS Public Datasets) for select industries like retail, energy, or manufacturing, to analyze changes in physical operational scale or geographical distribution.
  - **Audio Data:** Raw audio files from earnings conference calls will be sourced from commercial providers to enable the extraction of acoustic features that transcend textual content, such as emotional tone and vocal emphasis. Audio-visual materials from executive interviews and product launch events will be collected from corporate websites, financial media outlets, and conference archives.

3.1.2 Data Acquisition Process

The data acquisition strategy prioritizes systematic and robust collection:

- Institutional subscriptions to commercial database APIs will be the primary channel for obtaining standardized data.
- For public databases like EDGAR (U.S.) and SEDAR (Canada), automated download and parsing routines will be developed or existing Python libraries (e.g., sec-edgar-downloader, beautifulsoup4, requests) utilized.
- Web crawlers, designed to respect robots.txt protocols, will be employed for corporate websites and news portals.
- A comprehensive metadata inventory will be meticulously maintained, detailing the source, acquisition timestamp, and original format for every data point.

[4] https://www.crsp.org/research/crsp-historical-indexes

3.1.3 Preprocessing Pipeline

A critical preprocessing pipeline will transform raw, heterogeneous data into a clean, structured format suitable for sophisticated modeling. This phase employs advanced data processing techniques and standardized workflows.

- **Text Preprocessing**
  - Initial parsing and extraction will involve employing robust HTML/XML parsers (e.g., lxml) for EDGAR files and PDF parsing libraries (e.g., PyMuPDF/fitz , pdfminer.six) for PDF reports to precisely extract target sections or full-document text.
  - Subsequent cleaning will meticulously remove HTML tags, non-textual elements (e.g., chart placeholders), redundant paragraphs, and legal disclaimers, followed by standardization of text formats, including uniform UTF-8 encoding.
  - The cleaned text will then be segmented into sentences or paragraphs, forming the basic units for subsequent embedding.
  - For non-English texts, language detection using reliable tools (e.g., langdetect library) will be performed. High-quality machine translation APIs (e.g., Google Cloud Translation, DeepL API) will be applied to unify texts into English, where necessary, while diligently noting the original language; alternatively, multilingual versions may be retained for processing by compatible models.
  - An optional relevance filtering step may be explored for exceptionally long documents, such as annual reports. This could involve using lightweight embedding models (e.g., Sentence-Transformers) to compute semantic similarity between sentences and standard business description templates, thereby selecting the most pertinent sentence fragments to accommodate model input constraints while ensuring the integrity of critical information.
- **Structured Data Preprocessing**
  - Missing values will be addressed using appropriate imputation strategies (e.g., Imputation, or more sophisticated methods like K-Nearest Neighbors (KNN) imputation or model-based imputation).
  - Outliers will be identified and handled using techniques such as Z-score or Interquartile Range (IQR) methods. Duplicate records will be removed or consolidated.
  - Numerical features will be scaled to a uniform range (e.g., Min-Max Scaling, Standardization) as required by subsequent modeling steps.
  - New, meaningful features, such as financial ratios, will be engineered from the raw data where appropriate.
- **Visual Data Preprocessing**
  - Images will be standardized to uniform dimensions and formats (e.g., JPEG, PNG), with necessary enhancements or denoising applied.
  - Feature extraction will utilize pre-trained powerful vision models (e.g., CLIP, ViT, ResNet) to derive dense vector representations (embeddings) of images or to extract specific visual features like object labels or scene categories. For charts and

tables within documents, Optical Character Recognition (OCR) or specialized chart understanding models (conceptually similar to Microsoft Research Asia's FinVis) may be employed to extract structured information.

- **Audio Data Preprocessing**
  - High-quality Automatic Speech Recognition (ASR) services will be used to transcribe audio into text if transcripts are not already available from providers.
  - Standard acoustic features (e.g., MFCCs, pitch, energy, speech rate) will be extracted. Alternatively, pre-trained speech emotion recognition or speaker identification models may be used to derive higher-level features.

3.1.4 Cross-Cutting

Beyond modality-specific preparations, several cross-cutting procedures are fundamental to the integrity of the consolidated dataset.

Effective entity resolution and linking are paramount for integrating disparate data sources. Standard corporate identifiers (e.g., CUSIP, ISIN, SEDOL, LEI, Ticker Symbol) will serve as primary keys. An internal mapping table will be meticulously maintained to manage identifier changes stemming from corporate actions such as mergers, acquisitions, or name changes. Fuzzy matching algorithms and cross-referencing with resources like OpenCorporates will be employed to resolve entity ambiguities, ensuring each data point is unequivocally linked to a unique corporate entity.

Temporal alignment, an exceptionally critical phase, involves assigning the most precise event timestamp (reflecting the actual occurrence time of information) and as-of timestamp (indicating when information became publicly accessible) to every acquired data point, irrespective of its modality. For instance, an annual report's event time is its fiscal year-end, while its as-of time is its publication date . This painstaking alignment forms an indispensable foundation for subsequently addressing look-ahead bias.

Last but not least, rigorous data quality assurance protocols will be implemented across multiple tiers. This includes cross-validation of analogous information from different origins, continuous monitoring of data coverage and completeness, and the establishment of a rules-based engine to flag potential data errors or inconsistencies. Key data outputs, such as LLM-generated descriptions, will be subject to manual sampling audits. Comprehensive logging of all cleaning and transformation steps will ensure full process traceability.

Through above exacting data acquisition and preprocessing regimen, this research aims to construct an extensively covered, informationally rich, high-quality, and temporally aligned multimodal dataset. Such a dataset is conceived as the indispensable bedrock for the subsequent construction of a high-fidelity global financial business network.

3.2 Generating Standardized Company Descriptions (if applicable)

While direct extraction from primary textual sources remains the preferred methodology, practical exigencies-such as historical data lacunae, inconsistent quality of original texts, highly disparate formatting (particularly for non-English or scanned PDF reports), or excessive length precluding optimal processing by subsequent embedding models-necessitate a conditional and judicious approach. In such circumstances, a methodology leveraging SOTA LLMs will be employed to generate standardized, information-dense, and temporally precise corporate business descriptions, a technique shown to be viable for creating inputs for network construction.

This step is not intended to create information from zero but rather to function as a sophisticated tool for structured summarization and standardization, enhancing the quality and cross-sectional/temporal comparability of data fed into embedding models. Its implementation will adhere to the following meticulous and practicable process.

3.2.1 Applicability Criteria and Trigger Mechanism

The deployment of LLMs for description generation is strictly governed by pre-defined applicability criteria, underscoring the principle that LLMs function as sophisticated tools for structured summarization and standardization of extant information, rather than as original information generators. Such intervention is contemplated when:

- Exhaustive preprocessing fails to isolate a distinct and sufficiently detailed business description section (analogous to Item 1 in 10-K filings) from source documents.
- Extracted raw text significantly exceeds the optimal input length for embedding models, even after semantic similarity-based filtering (as detailed in the data preprocessing section) proves insufficient in producing concise yet comprehensive segments.
- Source documents are scanned PDFs yielding low-quality OCR output, thereby risking the introduction of substantial noise.
- For specific historical periods or geographic regions, descriptions from commercial data providers are known to be deficient or of questionable reliability.

3.2.2 Input Data Selection and Preparation

The integrity of LLM-generated descriptions hinges critically on the meticulous selection and preparation of input data.

Input Source Curation: Only publicly available, highly relevant original text segments predating or contemporaneous with the target fiscal period (t) will serve as LLM inputs. This may encompass:

- The most relevant paragraphs or sentence collections identified through semantic retrieval from complete annual reports.
- Pertinent content from the year's press releases and Management's Discussion & Analysis sections.
- Archival snapshots of the company's official website at the relevant time point (e.g., sourced via the Wayback Machine).

Strict Temporal Control: All textual fragments supplied to the LLM must possess timestamps no later than the target description date t. This is a crucial prerequisite for averting look-ahead bias, a significant concern when using models trained on data with recent knowledge cutoffs.

- **Strict Temporal Control:** All textual fragments supplied to the LLM must possess timestamps no later than the target description date t. This is a crucial prerequisite for averting look-ahead bias, a significant concern when using models trained on data with recent knowledge cutoffs.

- **Context Length Management:** Should the curated relevant text still exceed optimal LLM input lengths, strategies such as sliding windows or prioritization of key paragraphs will be employed to ensure ingestion of the most salient information.

- **Strict Temporal Control:** All textual fragments supplied to the LLM must possess timestamps no later than the target description date t. This is a crucial prerequisite for averting look-ahead bias, a significant concern when using models trained on data with recent knowledge cutoffs.

- **Context Length Management:** Should the curated relevant text still exceed optimal LLM input lengths, strategies such as sliding windows or prioritization of key paragraphs will be employed to ensure ingestion of the most salient information.

3.2.3 Model Selection and Configuration

The choice of LLM will be guided by demonstrated capabilities in instruction adherence, long-context comprehension, information synthesis, and, where pertinent, multilingual processing.

Model Considerations: Consideration will be given to leading-edge commercial LLMs and high-performance open-source alternatives (such as Llama 3, Mistral Large, or suitable finance-domain fine-tuned models, if available and validated).

Parameter Settings: Configuration parameters, notably a low temperature setting (e.g., 0.2-0.5), will be calibrated to foster deterministic and factually grounded outputs, thereby minimizing speculative or overly creative generation. Maximum output length constraints will also be applied to manage verbosity.

3.2.4 Refined Prompt Engineering

Prompt engineering is a linchpin for ensuring the quality and consistency of generated descriptions. Prompts will be meticulously crafted to include:

- **Role Definition:** e.g., ”You are a meticulous financial analyst tasked with composing a standardized description of [Company Name]’s business operations as of the end of fiscal year [YYYY], based on the provided source texts.”

- **Task Instruction:** e.g., ”Strictly using only the provided textual content, generate a concise (approximately 150-250 words), objective, and information-dense business description. The description should clearly cover (if mentioned in the text): (1) primary products, services, or business lines; (2) principal operating markets or geographic regions; (3) key customer segments or industries; (4) major strategic priorities or competitive advantages explicitly stated for that year; (5) significant business changes occurring during the year mentioned in the text (e.g., major acquisitions, divestitures, new product line launches).”

- **Constraints:** e.g., ”Absolutely prohibit the inclusion of any information not explicitly present in the input text. Forbid any form of inference, prediction, or subjective evaluation. Strictly adhere to the specified time point (end of fiscal year [YYYY]). If information on certain aspects is missing from the input text, omit that aspect in the description; do not fabricate.”

- **Output Format:** Requiring output in a coherent paragraph form, with a professional and neutral linguistic style.

- **Few-shot Learning:** The incorporation of 1-2 high-quality, compliant standardized descriptions as examples within the prompt may be employed to better guide the LLM's output style and content structure.

3.2.5 Cautious Exploration of Retrieval Augmented Generation

While Retrieval Augmented Generation (RAG) offers potential for enhancing LLM outputs by referencing external knowledge bases, particularly with highly fragmented inputs or for fact verification, its application will be approached with extreme caution. The inherent risk of introducing look-ahead bias via RAG is substantial. If RAG were to be employed, it would necessitate ensuring that the connected knowledge base is rigorously versioned by time, permitting the LLM to access only information contemporaneous with or preceding the target description point (t). This implies the use of vector databases or knowledge graphs supporting time-travel queries, a technically complex undertaking. Consequently, for the initial phases of this research, preference will be given to non-RAG approaches relying purely on curated textual inputs.

3.2.6 Quality Control and Iterative Optimization

Automated Preliminary Screening: Basic compliance checks (e.g., length, absence of proscribed terms like ”predict” or ”future”) will be performed. An independent embedding model may be used to calculate the semantic similarity between generated descriptions and input text fragments, flagging outputs with low input-relevance.

Manual Sampling Review (Crucial Step): The cornerstone of quality assurance will be rigorous manual review by researchers of a significant random sample (e.g., 5$\tilde{1}$0%) of LLM-generated descriptions against their original source texts. Evaluation criteria will encompass:

- **Factual Accuracy:** Is the generated content entirely grounded in the input text? Are there any fabrications or distortions?

- **Informational Completeness:** Are critical pieces of information regarding core business aspects from the input text omitted?

- **Temporal Consistency:** Is the specified time point strictly adhered to?

- **Conciseness and Standardization:** Does the output meet the required length and structural guidelines?

Iterative Feedback Loop: Systemic issues identified during manual audits (e.g., tendencies towards specific types of hallucination, misinterpretation of temporal constraints, biases in instruction following) will inform systematic adjustments to prompt design or input text selection strategies. This cycle of regeneration and re-evaluation will continue until the generated descriptions achieve a pre-defined quality threshold (e.g., $> 95\%$ manual audit pass rate).

3.3 Embedding Model Implementation and Selection

The selection of an optimal embedding model is paramount to ensure that the constructed business networks accurately capture the intricate economic relationships between enterprises. This research will systematically implement, evaluate, and select from state-of-the-art embedding models, following a meticulously planned process.

3.3.1 Candidate Model Identification and Preparation

A representative and high-performing suite of embedding models, spanning diverse technological approaches and deployment paradigms, will be curated for comparative analysis. This selection is designed to provide a comprehensive overview of the current embedding landscape.

- **Commercial Closed-Source API Models**

  This category includes prominent models known for their advanced capabilities and ease of access via APIs:

  - **OpenAI:** The text-embedding-3-small and text-embedding-3-large models are primary candidates, distinguished by their robust general semantic understanding and user-friendly API integration.
  - **Google Gemini text-embedding-004:** The Gemini text-embedding series represents Google's latest advancements in embedding technology, demonstrating strong performance across multiple benchmarks, and it's free for researchers by applying.
  - **Microsoft Azure AI:** The E5 embedding model family, exemplified by Azure's implementation of E5-large-v2, is notable for its leading performance on benchmarks such as the Massive Text Embedding Benchmark (MTEB).

  While these commercial offerings typically deliver high performance and obviate the need for local deployment infrastructure, practical considerations such as API invocation costs, rate limitations, and data privacy protocols will be carefully evaluated during their assessment.

- **Open-Source Models**

  To ensure flexibility, control, and facilitate reproducibility, a selection of leading open-source models will be included:

  - **Sentence Transformers (SBERT) Library:** Models from this library that excel in semantic similarity tasks and offer multilingual support are key considerations. Examples include all-mpnet-base-v2, a benchmark for English performance, and paraphrase-multilingual-mpnet-base-v2, known for its strong multilingual capabilities. The most current SOTA models available at the time of research execution will also be incorporated.
  - **Microsoft E5 Series (Local Deployment):** Variants such as intfloat/e5-large-v2, accessible via the Hugging Face Transformers library, offer finer-grained control over the model and its deployment.
  - **Potentially Fine-tuned Financial Models:** The landscape will be surveyed for publicly available embedding models specifically fine-tuned on financial corpora (e.g., embedding layers from finance-adapted BERT or RoBERTa variants) that exhibit superior performance. As an exploratory step, the feasibility of lightweight

fine-tuning of general-purpose models on proprietary financial text will also be considered.

Open-source models afford greater adaptability and control, with no direct API costs, but necessitate local computational resources, particularly GPUs for efficient inference, and may involve more intricate configuration.

- **Baseline Models** To contextualize the performance of advanced embedding models, several baselines will be established:
  - **Traditional Non-Embedding Method:** TF-IDF will serve as a representative of non-deep learning approaches.
  - **Early Embedding Methods:** Word2Vec (e.g., Skip-gram or CBOW), trained on the corpus of company descriptions to derive word vectors subsequently averaged or weighted to produce document vectors, will be implemented. For direct comparison, consideration will be given to replicating early embedding approaches similar to those utilized by B&M [5], such as their mentioned GPT-3 based embeddings or analogous techniques.

3.3.2 Embedding Vector Generation and Storage

The generation of embedding vectors from the prepared company descriptions is a critical operational step.

Input Data and Processing: Standardized, timestamp-aligned company description texts (either LLM-generated or directly extracted as per prior data processing stages) will form the input. For models requiring processing of long-form text (e.g., full annual reports rather than summaries), appropriate strategies such as chunking followed by averaging/max-pooling of embeddings, or the use of models specifically designed for long-text inputs, will be adopted.

- For API-based models, robust Python scripts will be developed to manage API authentication, batch requests, adherence to rate limits, and error retry logic.
- For open-source models, libraries such as Hugging Face transformers and sentence-transformers will be utilized for efficient inference on GPU-equipped servers. Batch sizes will be optimized to balance processing speed and memory consumption.

Output Specification: For each company at each relevant time point (e.g., annually), a fixed-dimension embedding vector will be generated (e.g., OpenAI v3 models yield 1536 or 3072 dimensions; E5-large yields 1024 dimensions).

Storage Solution: The resultant embedding vectors, along with their corresponding company identifiers (e.g., ISIN, PermID) and timestamps (e.g., year), will be stored efficiently in a suitable database (e.g., PostgreSQL with the pgvector extension for similarity search capabilities) or file formats optimized for large numerical datasets (e.g., Parquet, HDF5) to facilitate rapid retrieval and subsequent computational tasks.

Rigorous Quantitative Evaluation Framework: The cornerstone of model selection will be a multi-dimensional quantitative evaluation framework, meticulously designed for the specific task of financial network construction, thereby moving beyond generic NLP benchmarks. This evaluation will be conducted on a reserved validation dataset, diverse in time and geographic origin.

- **Dimension 1: Semantic Similarity Capturing**

  This dimension assesses the fidelity with which models capture nuanced semantic relationships.

  - **Correlation with Human Judgments:** A randomly selected set of company pairs (e.g., 200-300) will be rated for business similarity by domain experts (e.g., financial analysts or research team members) on a defined scale (e.g., 1-5). The Spearman Rank Correlation between these human ratings and the cosine similarities of embedding vectors generated by each model will be computed. Superior models are expected to exhibit higher correlation with human intuition.
  - **Known Relationship Recovery:** External databases (e.g., FactSet Revere Business Relationships, Bloomberg Supply Chain Data (SPLC), SDC Platinum M&A peer data) will serve as quasi-ground truth for known economic links.
  - **Metrics:** Precision@k / Recall@k (for each company, identifying its k-nearest neighbors based on embedding distance and calculating the proportion that are known competitors/suppliers/customers/M&A peers), Mean Average Precision (MAP) (evaluating the overall ranking quality of retrieving known related firms), and AUC (Area Under Curve) (assessing the model's ability to discriminate between "known related firm pairs" and "random non-related firm pairs" based on an embedding distance threshold).
  - **Expectation:** More effective models should position known related companies closer in the embedding space, leading to better performance on these metrics.

- **Dimension 2: Economic Structure Coherence**

  This evaluates the extent to which embeddings naturally reflect known economic structures.

  - **Industry/Country Clustering Performance:** Standard clustering algorithms (e.g., K-Means, DBSCAN) will be applied to company embedding vectors. The resulting clusters will be evaluated against known GICS industry classifications or company registration countries.
  - **Metrics:** Silhouette Score, Davies-Bouldin Index (DBI), and Adjusted Mutual Information (AMI).
  - **Expectation:** Embeddings that better capture economic realities should facilitate more coherent clustering of firms from similar industries or geographical regions.

- **Dimension 3: Downstream Task Performance**

  The practical utility of embeddings will be gauged by their performance as features in representative downstream financial applications (aligned with research object 4, e.g., M&A prediction or similarity-based pairs trading strategies). Simple predictive/trading models (e.g., Logistic Regression, XGBoost, or basic pairs trading rules) will be constructed using embedding vectors or derived similarity scores as core input features.

  - **Metrics:** For predictive tasks, AUC and F1-score will be used. For trading strategies, Sharpe Ratio, annualized return, and maximum drawdown will be assessed. Performance differences when using embeddings from various models will be compared.
  - **Expectation:** Embeddings that most accurately capture relevant economic linkages should translate to superior predictive power or profitability in these downstream applications.

- **Dimension 4: Computational Efficiency and Cost**

  Practical deployment considerations will be assessed. For local models, the average time to generate a single embedding vector will be recorded. For API models, API response times and associated costs will be logged to evaluate the feasibility of large-scale deployment.

3.3.3 Systematic Comparison and Model Selection

The final stage involves a comprehensive comparison and informed selection.

Execution: All evaluation metrics described above will be computed for every candidate model using the validation dataset. Results will be clearly tabulated and visualized for comparative analysis.

Statistical Testing: Statistical significance tests (e.g., paired t-tests or bootstrap methods) will be applied to differences in key performance indicators (such as downstream task performance or AUC in known relationship recovery) to ascertain the reliability of observed performance variations.

Sensitivity Analysis: A brief examination of the impact of different similarity measures (e.g., cosine similarity versus Euclidean distance) and critical parameters (e.g., the value of k in k-NN) on the results will be conducted.

Final Decision-Making: The ultimate model selection will be based on a holistic assessment of performance across all dimensions, with particular emphasis on Dimensions 1, 2, and 3, alongside considerations of computational cost and scalability. Preference will be given to models demonstrating superior performance on tasks directly relevant to financial network construction, especially known relationship recovery and downstream task efficacy. The possibility of employing a model ensemble will not be precluded. For instance, if one model excels at capturing competitive relationships while another is superior for supply chain links, strategies for fusing these embeddings (e.g., simple or weighted averaging) may be explored.

3.4 Multimodal Fusion Framework

The unimodal reliance on textual data, while capable of capturing the semantic nuances of corporate business descriptions, demonstrably overlooks a wealth of supplementary signals embedded within visual elements (e.g., product aesthetics, brand insignia, financial charts), acoustic characteristics (e.g., managerial sentiment and confidence conveyed during earnings calls), and structured data (e.g., financial health indicators, ESG performance metrics). These non-textual cues are frequently indispensable for a holistic comprehension of a firm's market positioning, competitive advantages, latent risks, and, critically, its authentic economic interconnections with other entities. The growing focus on multimodal data in financial NLP underscores this shift, recognizing that a comprehensive analysis requires integrating these diverse information streams.

Consequently, this study proposes the development of a Multimodal Joint Embedding framework. The central objective is to project enterprise-related information, derived from these disparate modalities, into a shared, low-dimensional semantic space. Within this unified representational space, the proximity of resultant vectors is engineered to reflect comprehensive economic relatedness with a precision that transcends the analytical scope afforded by any single modality.

3.4.1 Modality-Specific Feature Extraction

Prior to multimodal fusion, robust feature extraction is performed for each data modality to generate vector representations amenable to integration.

Textual Modality: Standardized corporate business descriptions, key excerpts from annual reports, earnings call transcripts, and relevant news summaries serve as inputs. The optimal text embedding model, as determined in the ”Embedding Model Implementation and Selection” phase (e.g., text-embedding-3-large, Gecko, or a fine-tuned Sentence-BERT variant), will be employed.

This process yields a high-quality, fixed-dimension text embedding vector, v_{text}, for each company at specific time points.

Visual Modality: Inputs include product images, corporate logos, informative charts from annual reports or websites (requiring preprocessing for identification, e.g., market share diagrams), and potentially satellite imagery of physical assets (e.g., factories, retail locations). Powerful pre-trained Vision-Language Models, such as CLIP (Contrastive Language–Image Pre-training) utilizing its image encoder (ViT or ResNet variants), or dedicated image feature extractors like EfficientNet, will be leveraged. CLIP offers an intrinsic advantage as its image and text embeddings are co-trained for alignment, facilitating subsequent fusion. For chart-based imagery, initial processing via OCR or specialized chart interpretation models (e.g., approaches similar to FinVis may be necessary to extract structured information or generate descriptive text prior to embedding [46].

This results in one or more image embedding vectors, v_{image}, per company, which, if multiple, can be aggregated into a single representative vector using techniques like average pooling, max pooling, or an attention-based mechanism.

Auditory Modality: Audio segments derived from investor conference calls, executive interviews, and similar vocal recordings will serve as input for this modality. The primary analytical focus here is the extraction of affective and paralinguistic cues that extend beyond the literal textual content. Two principal strategies are envisaged for feature extraction:

- The utilization of pre-trained SER models, potentially based on architectures like Wav2Vec 2.0 or HuBERT, to directly derive emotion classification probabilities or dedicated emotion embedding vectors.

- The extraction of classic acoustic feature sets (e.g., Mel-Frequency Cepstral Coefficients (MFCCs), pitch contours, energy levels, zero-crossing rates), which would then be fed into a compact neural network (such as an LSTM or GRU) to learn temporal patterns and generate a consolidated voice embedding.

The output of this stage is an audio/vocal embedding vector, v_{audio}, designed to represent managerial sentiment, confidence, or characteristic speech style for each relevant recording.

Structured Modality: A curated set of key financial ratios (indicative of profitability, leverage, liquidity), market capitalization figures, volatility metrics, ESG scores, and industry classification codes will constitute the inputs for the structured data modality. Given the inherent heterogeneity in scale, dimensionality, and type (numerical and categorical) of these data:

- Numerical features will undergo rigorous standardization, for example, through Z-score normalization.
- Categorical features will be appropriately transformed using techniques such as one-hot encoding or target encoding.
- The processed structured features will then be projected into a lower-dimensional embedding space. This is typically achieved using a simple Multi-Layer Perceptron (MLP) or an Autoencoder, a process which aids in capturing potential non-linear relationships among the features and rendering the structured information compatible with the vector spaces of other modalities.

This procedure will yield a structured data embedding vector, v_{struct}, for each company at specific, relevant time points.

3.4.2 Fusion Architecture Design and Implementation

The efficacy of the overarching framework critically depends on the design and implementation of robust fusion mechanisms. Several viable and innovative strategies will be systematically explored and comparatively evaluated.

A foundational approach, serving as a baseline, is Concatenation-based Simple Fusion. This method involves the direct concatenation of all individual modal embedding vectors. Missing modalities will be addressed through appropriate imputation, for instance, using zero-vectors or modality-specific mean vectors. The resultant higher-dimensional composite vector can then be optionally passed through one or more fully connected layers, incorporating non-linear activation functions (e.g., ReLU), for dimensionality reduction and to facilitate feature interaction, ultimately yielding the final fused embedding, v_{fused}. While this strategy benefits from simplicity of implementation and relatively modest computational overhead, its primary limitation is the absence of explicit modeling for complex inter-modal interactions, which may curtail its capacity to fully exploit potential multimodal synergies.

A more sophisticated while potentially more powerful strategy is Attention-based Weighted Fusion. This approach leverages either self-attention or cross-modal attention mechanisms. Self-attention mechanisms can treat embeddings from different modalities as distinct elements within a sequence, applying Transformer encoder layers. This allows the model to learn the differential importance (weights) of various modalities for the final aggregated representation and to capture nuanced interactions among them. Alternatively, dedicated cross-modal attention modules can be designed. For example, the text embedding v_{text} could act as a query to attend to relevant information within the image embedding v_{image} and the audio embedding v_{audio} (which would serve as keys and values), and vice versa for other modal pairings. The principal advantage of attention-based fusion is its ability to dynamically adjust the contribution of each modality based on the specific context, thereby more effectively capturing intricate cross-modal dependencies. However, this enhanced capability comes at the cost of increased model complexity and computational demand, often necessitating larger datasets for effective training.

The primary exploratory direction within this research, however, is Contrastive Learning for Joint Embedding. This strategy is designed to learn a shared embedding space where economically related enterprises—irrespective of the modality through which they are represented—are positioned in closer proximity, while unrelated enterprises are mapped further apart. The architecture typically involves multiple modality-specific encoders, which can be pre-trained models augmented with trainable projector layers, responsible for mapping inputs from various modalities into this common, unified embedding space.

The construction of **Positive Pairs** is a critical element of this approach. These pairs can be formed in two primary ways:

- **Cross-company association:** Pairing the text embedding of Company A with the text embedding of Company B, where A and B are known to be competitors, part of a supplier-customer dyad, or counterparts in a recent M&A transaction.
- **Intra-company cross-modal alignment:** Ppairing the text embedding of Company A with the product image embedding of Company A, or the text embedding of Company A with the earnings call audio sentiment embedding of Company A. This type of pairing is instrumental in achieving robust alignment of representations across different modalities for the same underlying entity.

Negative Pairs, conversely, might consist of the embedding of Company A paired with the embedding of Company C, where Company C is a firm randomly sampled and demonstrably different from A in terms of industry, geography, or other relevant economic characteristics. The design of sophisticated negative sampling strategies is vital to mitigate the risk of ”false negatives” (i.e., inadvertently pairing entities that do, in fact, share a latent economic relationship).

The Loss Function employed in such framework is typically InfoNCE (Noise Contrastive Estimation) or one of its established variants, such as the contrastive losses utilized in SimCLR or MoCo. The overarching objective is to maximize the similarity (e.g., cosine similarity) of positive pairs while simultaneously minimizing the similarity of negative pairs.

The principal merit of this contrastive learning approach is its direct optimization towards the goal of ”relationship discovery,” which should, in theory, yield representations that are more genuinely reflective of underlying economic interconnectedness. Furthermore, it inherently accommodates scenarios with missing modalities, as the contrastive loss can generally be computed provided that at least one modality is present for each entity in a given pair. Nevertheless, the performance of contrastive learning frameworks is acutely sensitive to the quality of positive and negative pair construction, demanding either high-quality prior knowledge (in the form of known relationships) or highly effective self-supervised signals. The training process for such models can also be computationally intensive.

3.4.3 Training Strategy and Handling Missing Modalities

The training strategy will be carefully adapted to the chosen fusion strategy. Should contrastive learning (Strategy 3) be the primary focus, the training will predominantly be unsupervised or self-supervised, relying heavily on the meticulously curated positive and negative sample pairs.

For concatenation-based or attention-based fusion approaches (Strategies 1 or 2), particularly if the objective involves the direct prediction of a specific type of relationship or attribute, some degree of supervised data might be necessary for fine-tuning the model. Across all strategies, standard best practices in model training—including the use of appropriate optimizers (e.g., AdamW), learning rate scheduling policies, and robust regularization techniques (e.g., Dropout, Weight Decay)—will be rigorously employed to prevent overfitting and ensure generalization.

Addressing the issue of missing modalities, an almost inevitable characteristic of real-world financial datasets, is of critical importance. During the training phase, a technique known as modality dropout will be implemented. This involves randomly masking or omitting certain modal inputs during training iterations, thereby compelling the model to learn inferentially from the remaining available modalities and enhancing its overall robustness to incomplete data. During the inference phase, several tactics can be employed:

- For modalities that are entirely absent for a given data point, pre-defined zero vectors or mean embedding vectors (derived from the training set average for that particular modality) can be inputted as placeholders.

- If the fusion architecture is designed to accommodate it (as is often the case with certain attention mechanisms), the model can be engineered to automatically disregard or assign zero weight to absent modal inputs.

- Within contrastive learning frameworks, the presence of at least one modality for each entity in a pair is generally sufficient to compute its embedding and subsequently use it for similarity comparisons, offering a degree of inherent resilience to missing data.

3.4.4 Output and Application

The ultimate output of this multimodal fusion framework, for each company at every relevant time point t, will be a unified, low-dimensional multimodal fused embedding vector, denoted as v_{fused_t}. This vector serves as the coordinate of that company's comprehensive economic profile within the shared semantic space at that specific juncture. These v_{fused_t} vectors will then constitute the core input for the subsequent ”Business Network Construction” phase of the research. The strength and nature of inter-company relationships will be quantified by computing the distance or similarity (e.g., cosine similarity) between their respective v_{fused_t} vectors. This quantitative measure of relatedness will, in turn, enable the construction of an enhanced, nuanced, and dynamic multimodal global business network.

3.5 Business Network Construction

Upon the generation of company embedding vectors, encompassing both text-based baseline embeddings (v_{text}) and enhanced multimodal-fused embeddings (v_{fused_t}), the research will transition to the practical construction of business networks. The core task at this stage is to translate distances or similarities within the high-dimensional, continuous embedding space into discrete or weighted graph structures. In these graphs, nodes represent companies, and edges signify their latent economic relationships. The construction process will involve a systematic exploration and comparison of various methodologies, alongside an endeavor to identify and differentiate distinct types of economic connections.

3.5.1 Quantifying Relationships based on Embedding Similarity

The fundamental premise, extending the work of Hoberg and Müller, is that companies in close proximity within a semantic embedding space are more likely to share similar or related business models, product markets, risk exposures, or strategic priorities [5, 19, 20]. This study will primarily employ Cosine Similarity to quantify the resemblance between the embedding vectors $emb_{i,t}$ and $emb_{j,t}$ of company i and company j at a specific time point t. Cosine similarity, which focuses on the orientation rather than the absolute magnitude of vectors, is particularly well-suited for high-dimensional spaces. The formula for calculation is: $Sim(i, j, t) = \frac{emb_{i,t} \cdot emb_{j,t}}{||emb_{i,t}||||emb_{j,t}||}$. Separate similarity matrices will be computed based on both v_{text} and v_{fused_t} to allow for a comprehensive comparison of how different embedding approaches influence network structure.

3.5.2 Network Formation Strategies: From Similarity to Edges

Translating continuous similarity scores into network connections (edges) can be achieved through various strategies. This research will systematically implement and compare the following primary methods to determine the most effective approach for reflecting true economic relationships and optimizing predictive performance:

Thresholding This method involves setting a dynamically determined similarity threshold τ. An undirected edge is established between company i and company j if $Sim(i, j, t) > \tau$.

- **Parameter Selection:**The choice of threshold τ is critical. Multiple approaches will be explored to determine the optimal value: (a) based on intrinsic evaluation metrics, selecting the threshold that maximizes the network's match (e.g., F1-score) with known relationships from external datasets (e.g., FactSet Revere competition/supply chain data); (b) based on network structural characteristics, choosing a threshold that produces network densities or average degrees consistent with expectations from economic literature; and (c) based on extrinsic evaluation performance, selecting the threshold through cross-validation that yields the best predictive performance in downstream tasks (e.g., earnings prediction).

- **Potential Drawbacks:**This method is highly sensitive to the chosen threshold, which can lead to significant variations in network density.

k-Nearest Neighbors (k-NN) For each company i, this method identifies the k most similar other companies in the embedding space and establishes connections.

- **Variants and Symmetrization:**Connections can be unidirectional (j is a k-NN of i) or symmetrized. Symmetrized variants include: (a) Mutual k-NN, where a connection is established only if i is a k-NN of j and j is also a k-NN of i; and (b) OR Connection, where a connection is formed if either i is a k-NN of j or j is a k-NN of i. This research will primarily employ symmetrized methods to construct undirected graphs.

- **Parameter Selection:**The selection of k is equally crucial. Optimization will follow strategies similar to thresholding: (a) based on intrinsic evaluation (matching known relationships); (b) based on downstream task performance; and (c) referencing commonly used k values in relevant literature.

- **Potential Drawbacks:**While k-NN offers better control over network density and is less sensitive to absolute similarity values, it may inadvertently overlook important connections that fall below the k-NN cutoff but still represent high similarity.

Weighted Network This strategy directly assigns the calculated $Sim(i, j, t)$ as the weight $w(i, j, t)$ of the edge between company i and company j. Typically, a lower similarity threshold (e.g., > 0 or a baseline value) is applied to filter out highly irrelevant connections and manage network scale.

- **Application:**Weighted networks preserve information about the strength of relationships and can be directly utilized by network analysis algorithms that require weighted inputs (e.g., weighted centrality measures, community detection algorithms like the Louvain method).

- **Potential Drawbacks:**While retaining the most complete information, weighted networks inherently introduce higher analytical complexity.

This study will construct both binary (unweighted) networks using thresholding and k-NN, and weighted networks, systematically comparing their performance across various tasks in subsequent evaluation phases.

3.5.3 Dynamic Network Construction

Given the dynamic nature of the business environment, this research will construct time-series networks. The process will involve:

1. Determining time windows (e.g., annually or quarterly), informed by data availability (e.g., annual report release frequencies).

2. For each time window t, company embeddings ($emb_{i,t}$) will be generated using information available up to or at that time point.

3. The inter-company similarity matrix (Sim_t) will be computed for the specific time point t.

4. The chosen network formation strategy (e.g., threshold τ_t or k-NN k_t, with parameters potentially optimized dynamically or held constant across time) will then be applied to construct a network snapshot $G_t = (V, E_t, W_t)$, where V denotes the set of nodes, E_t the set of edges, and W_t optional edge weights.

5. This process will yield a sequence of network snapshots $G_{t1}, G_{t2}, ..., G_T$, which can be subsequently analyzed for dynamic characteristics such as network structure evolution, connection persistence, and changes over time.

3.5.4 Exploring Relationship Type Differentiation

Existing similarity-based networks generally lack the capability to distinguish the specific economic meaning of connections (e.g., competitor, supplier, customer, or other associations). This research will explore several feasible approaches to differentiate relationship types, aiming to infuse the network with richer economic semantics.

- **Supervised Learning-Based Classifiers**

  - **Data Preparation:**Partial company pair datasets with relationship type labels will be generated using external databases (e.g., FactSet Revere, Bloomberg SPLC) or through keyword matching in annual reports (e.g., explicit mentions of ”competitor,” ”supplier,” ”partner”). The potential for noise and incompleteness in these proxy labels is acknowledged.
  - **Feature Engineering:**For a given pair of companies (i, j), input features will include: (a) their respective embedding vectors $emb_{i,t}, emb_{j,t}$; and (b) composite features derived from the embedding vectors, such as element-wise differences $|emb_{i,t} - emb_{j,t}|$ or products $emb_{i,t} \cdot emb_{j,t}$, which have been shown to capture relational information.
  - **Model Training:**A multi-class classifier (e.g., Logistic Regression, Support Vector Machine, or a small neural network/MLP) will be trained on the labeled data to predict the most probable relationship type for a given company pair (e.g., competition, supply chain, complementarity/collaboration, unrelated).
  - **Application:**The trained classifier will then be applied to all (or high-similarity) connections within the network to assign probabilistic relationship type labels.

- **LLMs Zero/Few-shot Inference**

  This method involves crafting a prompt for a pair of companies (i, j) that includes their standardized text business descriptions (not embedding vectors) and any relevant contextual information (e.g., industry, country).

  - **Prompt Design (Example):**”Analyze the following business descriptions for two companies and determine their primary economic relationship (options: direct competitor, potential competitor, supplier-customer relationship, strategic partner, largely unrelated). Briefly explain your reasoning.A (Industry: XX, Country: YY): [Description of Company A]...B (Industry: ZZ, Country: WW): [Description of Company B]...”
  - **Feasibility:**While computationally expensive and likely unsuitable for classifying all millions or billions of potential network connections, this method can be strategically employed for: (a) validating the results of supervised classifiers; (b) in-depth analysis of high-priority company pairs (e.g., network hub nodes, companies involved in specific events); and (c) generating high-quality labeled data to train supervised classifiers.

- **Leveraging Multimodal Feature Cues**

  This approach is predicated on the hypothesis that different types of economic relationships may manifest distinct signal strengths across specific data modalities. For instance, competitors might exhibit greater similarity in product imagery, while supply chain partners might show stronger correlations in geographical location (if using satellite imagery) or specific technical terminology (in textual data).

  - **Implementation Exploration:**(a) Within supervised classifiers, the individual contribution of unimodal embeddings (textual, visual, etc.) to predicting specific relationship types will be assessed; (b) If attention-based fusion models are used, an analysis of cross-modal attention weights will explore which modal interactions the model prioritizes when determining relationships; and (c) Structural data, such as shared patents or common board directorships, will be integrated to further aid in relationship type inference.

3.5.5 Structured Network Data Output

The final output of this stage will be structured network data, stored in formats amenable to subsequent graph analysis (e.g., using NetworkX or ‘graph-tool‘ libraries). This will include:

- **Node List:**Containing company identifiers and associated attributes (e.g., name, industry, country, embedding vectors for time t).
- **Edge List:**Specifying connected company pairs (i, j), their corresponding time point t, edge weights $w(i, j, t)$ (if applicable), and potentially predicted relationship type labels with associated confidence scores.
- **Metadata:**Documenting the specific methods used for network construction, including the chosen embedding models, similarity metrics, network formation strategies, and parameter settings.

These structured network data will then serve as the input for the subsequent Evaluation Framework, enabling a comprehensive assessment of their quality and economic value through both intrinsic and extrinsic metrics.

3.6 Evaluation Framework

To ascertain the scientific rigor and practical utility of the enhanced global financial business network constructed in this research, a multi-dimensional and multi-layered evaluation framework will be designed and implemented. This framework is conceived not only to assess the structural soundness and relational accuracy of the network itself (intrinsic evaluation) but, more critically, to test its efficacy in addressing substantive financial problems (extrinsic evaluation). Through stringent benchmarking and comprehensive ablation studies, the innovative contributions of the proposed methodology will be rigorously validated.

3.6.1 Intrinsic Evaluation: Assessing Network Quality and Accuracy

The intrinsic evaluation will directly measure the constructed network's structural properties and its fidelity to known, real-world economic relationships.

- **Network Topology Analysis**

  A foundational analysis of the network's topological characteristics will be conducted using standard graph analysis libraries such as NetworkX or igraph in Python. Key network metrics will be computed to assess the network's structural integrity and economic plausibility. These metrics include the **degree distribution**, examined for scale-free properties indicative of hub-and-spoke structures common in economic systems; the **clustering coefficient**, to measure local interconnectedness reflective of industry or regional agglomeration; and the **average path length**, to evaluate the efficiency of information propagation across the network. Furthermore, community detection algorithms, such as the Louvain method, will be employed to identify modular structures, which will then be validated against known industry, regional, or supply chain clusters. The analysis of various **centrality measures** (e.g., degree, betweenness, eigenvector centrality) will serve to identify systemically important firms or ”keystone” nodes within the network. Comparing these topological features against established economic theories (e.g., core-periphery structures, industry clustering) and empirical observations will provide a robust assessment of the network's economic intuition. This analysis will also systematically compare the structural differences arising from networks built using different methodologies (e.g., text-only vs. multimodal embeddings, different embedding models).

- **Validation Against Ground-Truth Relationships**

  The most direct test of the network's verisimilitude involves benchmarking its connections against established ground-truth databases. To this end, recognized third-party data sources will be procured and processed.

  - **Competitive Relationships:**Data will be sourced from FactSet Revere Business Relationships (Competitors), Bloomberg Industry Classification System (BICS) defined peers, and entities identified as competitors by major market regulators.
  - **Supply Chain Relationships:**Data will be drawn from FactSet Revere Supply Chain Relationships, Bloomberg Supply Chain data (SPLC), and, where structurally extractable, company-disclosed lists of major customers and suppliers.
  - **Other Associations:**Links such as target-acquirer pairs in M&A transactions and strategic alliance partners will also be incorporated.

  For each pair of companies with a known relationship, an assessment will be made to determine if a corresponding link exists (for unweighted networks) or if a high similarity score is present (for weighted networks). The performance will be quantified using standard metrics such as **Precision**, **Recall**, and the **F1-Score**. For weighted networks, Receiver Operating Characteristic (ROC) curves will be plotted and the **Area Under the Curve (AUC)** calculated to evaluate the network's predictive power in distinguishing true relationships from random non-relations. This validation is one of the most critical indicators of the constructed network's quality.

- **In-depth Case Studies**

  To complement the quantitative metrics, which may overlook fine-grained details, in-depth qualitative case studies will be conducted. Several representative industries (e.g., electric vehicles, semiconductors, biotechnology) or specific economic events (e.g., the initial supply chain shocks of the COVID-19 pandemic in 2020, a major cross-border merger) will be selected. The relevant sub-networks will be extracted and visualized. Through expert analysis, informed by industry knowledge and detailed review of contemporaneous literature and news, a manual assessment will be performed to judge whether the sub-network's structure—including key players, core connections, and cluster distributions—accurately and reasonably reflects the true business landscape and dynamics of that specific scenario. This qualitative validation enhances the credibility and interpretability of the overall findings.

3.6.2 Extrinsic Evaluation: Testing Utility in Downstream Financial Tasks

The extrinsic evaluation will gauge the network's effectiveness and economic significance by applying it to solve concrete financial prediction and analysis tasks, with a focus on the application scenarios outlined in the research objectives.

- **Lead-Lag Effects in Global Stock Returns**

  The network's ability to capture information flow will be tested by constructing investment portfolios based on network linkage strength or centrality metrics. For instance, strategies will be developed to go long on leading firms and short on lagging firms within highly connected pairs, or to form portfolios long on high-centrality firms and short on low-centrality firms. Additionally, network-based factors (e.g., average returns of neighboring companies, centrality factors) will be constructed and tested for significant alpha within standard asset pricing models (e.g., Fama-French models). The statistical significance and economic meaning of lead-lag relationships will be further examined using Granger causality tests or Vector Autoregression (VAR) models. The primary evaluation metrics will be the **Sharpe Ratio**, **Information Ratio**, and **annualized alpha** of the constructed portfolios, alongside the statistical significance of the regression and causality test results.

- **M&A Event Prediction**

  To assess the network's predictive power for M&A events, network-derived features—such as the embedding similarity between firm pairs, number of common neighbors, network distance, and the centrality of target/acquirer firms—will be integrated into standard M&A prediction models (e.g., Logistic Regression, SVM, XGBoost), which typically include firm financials, size, and industry as control variables. The marginal contribution of these network features will be evaluated by the significant uplift in the model's **AUC** and **F1-score**, as well as the **feature importance** ranking of the network-derived variables within the predictive model.

- **ESG Spillover Analysis**

  The network's capacity to model contagion effects will be evaluated by analyzing the spillover of ESG performance. Spatial econometric models, such as the Spatial Autoregressive (SAR) or Spatial Durbin Model (SDM), will be employed. In these models, a company's ESG performance (e.g., change in ESG score, probability of a major ESG controversy) will be regressed on its own lagged values, other control variables, and a crucial spatial/network lag term ($W \cdot Y$). Here, W is a spatial weight matrix constructed based on the business network's connections (binary or weighted by similarity), and Y is the vector of ESG performance of neighboring firms. The key evaluation metric will be the statistical significance, sign, and magnitude of the estimated spatial autoregressive coefficient (ρ or λ), which indicates whether and how ESG performance propagates through the business network.

- **Global Supply Chain Resilience Assessment**

  The network's utility for risk analysis will be tested through simulations of supply chain disruptions. Various shock scenarios will be designed, such as the removal of key nodes (simulating a major supplier bankruptcy or a geopolitically induced disruption) or a reduction in edge weights (simulating increased transportation costs or tariff barriers). The network's resilience will be quantified by measuring structural changes before and after the shock, such as shifts in the size of the largest connected component, changes in average path length, or declines in network efficiency, allowing for the identification of cascading failure pathways. Where feasible, if firm-level operational data (e.g., inventory, sales) are available, these structural changes will be linked to economic indicators to assess potential economic losses. The evaluation will be based on network robustness metrics (e.g., resilience indices), the identification of critical paths, and, if possible, the magnitude of simulated economic impacts.

3.6.3 Validation through Benchmarking and Ablation

To rigorously validate the methodological contributions of this research, a comprehensive comparison against established benchmarks and a series of ablation studies will be conducted.

- **Benchmark Comparison**

  The performance of the proposed networks will be strictly compared against several well-chosen benchmarks across all intrinsic and extrinsic evaluation metrics:

  - **Benchmark I-Traditional Industry Classification Networks:** Networks constructed based on standard industry classifications like GICS, NAICS, or SIC, where firms in the same sub-industry are considered fully connected.
  - **Benchmark II-Stock Return Correlation Networks:** Weighted networks built from the rolling correlation or covariance of firm stock returns.
  - **Benchmark III-Key Literature Replication:** A faithful replication of the core methodology from representative studies, using their recommended embedding models and processing pipelines.
  - **Benchmark IV-Simplified Text-Embedding Networks:** Networks built using the same textual data sources but employing simpler or earlier embedding techniques (e.g., TF-IDF with Cosine Similarity, averaged Word2Vec vectors, or a basic SentenceBERT model).

  Performance differences between the proposed methods (both text-only and multimodal versions) and these benchmarks will be systematically documented and tested for statistical significance.

- **Ablation Studies**

  To isolate and quantify the independent contribution of each innovative component of the methodology, targeted ablation studies will be performed:

  - **Impact of Embedding Model:**The network constructed with the final selected SOTA embedding model will be compared against one built with a baseline embedding model (e.g., from Benchmark 4) to quantify the performance gain attributable to the advanced model.
  - **Impact of Multimodal Fusion:**The network built using the full multimodal fusion framework will be compared against a network constructed using only the textual modality (with the same advanced text embedding model). This comparison, particularly in tasks where multimodal information is hypothesized to be valuable (e.g., those involving product visual similarity or managerial sentiment), will reveal the marginal benefit of incorporating non-textual data. Further comparisons against single-modality networks (e.g., image-only) may also be conducted.
  - **Impact of Look-ahead Bias Control:**For historical backtesting tasks (e.g., lead-lag strategies, historical M&A prediction), the performance of networks built using the proposed bias control strategies (e.g., context-aware masking) will be compared against those built with no or simplistic masking. This will be crucial for observing whether inflated predictive performance has been effectively mitigated.

By synergistically combining intrinsic evaluation, extrinsic evaluation, benchmark comparisons, and ablation studies, this framework will facilitate a comprehensive, objective, and in-depth assessment of the quality, value, and innovation of the enhanced global financial business network, providing robust empirical support for the research conclusions.

3.7 Addressing Look-ahead Bias

Look-ahead bias presents a core challenge and a potential methodological pitfall when employing pre-trained LLMs or embedding models for the analysis of historical financial time-series data. These models, by virtue of their training on vast corpora, may have **memorized** events, knowledge, or linguistic patterns that post-date their stated knowledge cutoff. When applied to analyze historical data preceding this cutoff, they can inadvertently leverage this future information, leading to distorted representations of historical relationships and artificially inflated predictive performance in backtesting scenarios. Failure to adequately address this issue would severely compromise the veracity of the historical business networks constructed in this research and the reliability of any derivative analyses.

Consequently, this study will adopt a systematic, multi-pronged strategy to proactively identify, quantify, and mitigate look-ahead bias to the greatest extent possible, thereby ensuring the temporal consistency of the constructed networks.

- **Strict Temporal Discipline in Input Data**

  This serves as the primary defense against look-ahead bias. The core principle is that for any analysis pertaining to a specific time point t, only information that was publicly available at or before t can be used. This will be implemented through several measures:

  - **Precise Timestamping:**During the ”Data Acquisition and Preprocessing” stage, all data sources—including annual report release dates, conference call dates, news timestamps, financial data cutoff dates, and image/video metadata—will be assigned the most precise and conservative ”as-of” timestamp available. Particular attention will be paid to the publication lags of financial reports and macroeconomic indicators, using their actual release dates rather than just the reporting period end dates.
  - **Information Filtering:**When providing context for LLM-based description generation or when inputting text and multimodal data into embedding models, any information with a timestamp later than the target analysis point t will be strictly filtered out.

- **Targeted Prompt Engineering for LLM Generation (if applicable)**

  Should LLMs be employed for generating historical business descriptions, the prompts will be carefully engineered to guide the model to operate within a ”past-tense” and information-constrained framework.

  - **Explicit Temporal Constraints:**Prompts will include clear instructions such as, ”Based on information available as of [Date], describe the primary business of [Company Name] at that time...”
  - **Contextual Limitation:**The contextual material provided to the LLM will strictly adhere to the timestamping principles outlined above.

  - **Avoidance of Future Implications:**The LLM will be instructed to avoid language that implies future developments or uses terms and concepts that only emerged after the specified time point.

- **Advanced Named Entity Recognition for Identifying Potential Future Information Carriers**

  This step aims to automatically identify specific entities within the text that could potentially leak future information, providing precise targets for subsequent masking or anonymization. This approach is more reliable and comprehensive than simple keyword matching.

  - **Model Selection:**A pre-trained NER model with demonstrated high performance in the financial domain (e.g., based on RoBERTa or XLM-R, possibly fine-tuned on financial corpora for enhanced accuracy on domain-specific entities like product names, technological terms, and M&A events) will be utilized. Given the global and multilingual nature of this research, preference will be given to NER models with strong cross-lingual capabilities.
  - **Entity Types:**The focus will be on identifying entities such as: (1) Future product or service names not yet publicly announced at time t; (2) Future events like unannounced M&A deals, divestitures, key personnel changes, or significant litigation outcomes; (3) Unreleased financial data or forecasts beyond the known reporting period; (4) Company or brand names that did not exist or were not relevant at time t.
  - **Precision Requirement:**Given the financial context, the NER model must exhibit high precision, prioritizing the avoidance of incorrectly masking crucial historical information over the risk of missing some future-leaking entities, which can be addressed through other means.

- **Context-Aware Masking or Anonymization**

  Once potential future information carriers have been identified, they must be handled in a way that prevents the embedding model from leveraging its ”memory” of future knowledge, while maximally preserving the original text’s semantic context, which is vital for embedding quality. Simple deletion of entities is not viable as it would disrupt sentence structure and meaning.

  - **Context-Aware Masking:**Identified future entities will be replaced with meaningful, typed placeholders. For example, a yet-to-be-released product name could be replaced with ‘[FUTURE_PRODUCT_NAME]‘, an unannounced M&A target with ‘[UNANNOUNCED_M&A_TARGET]‘, and a future CEO’s name with ‘[FUTURE_CEO]‘. This method preserves sentence structure and provides type information, compelling the embedding model to rely on the surrounding context for understanding.
  - **Entity Anonymization:**As a comparative or alternative strategy, identified future entities can be replaced with unique but semantically void identifiers (e.g., ‘[ENTITY_123]‘ or a random string). This may more thoroughly eliminate the memory effect of the entity itself but could result in some loss of type information. The effectiveness of both strategies will be compared.

    - **Implementation Tools:**Scripts will be developed to automate the text replacement process, integrating the output of the NER model to generate a ”sanitized” version of the text for embedding.

- **Systematic Evaluation of Bias Impact and Strategy Validation**

  It is insufficient to merely implement mitigation strategies; their effectiveness must be quantitatively assessed. This involves quantifying the potential impact of look-ahead bias and validating the chosen mitigation approaches.

    - **Comparative Experimental Design:**
        - **Baseline Group:**Embeddings and networks generated using original, unprocessed data.
        - **Treatment Group 1:**Embeddings and networks generated from data processed with our optimal context-aware masking strategy.
        - **Treatment Group 2 (optional):**Embeddings and networks from data processed via entity anonymization.
        - **Treatment Group 3 (optional):**A comparison group using simpler masking methods proposed in other literature (e.g., simple NER-based masking as potentially used in B&M) [5].
    - **Evaluation Dimensions:**
        - **Changes in Historical Backtesting Performance:**The historical backtesting performance of downstream tasks (e.g., trading strategies based on lead-lag effects) will be compared across groups. A decrease in excess returns (alpha) in the treatment groups compared to the baseline would be interpreted as evidence of successful bias removal, indicating that the baseline’s high returns were partly spurious. Changes in risk-adjusted return measures like the Sharpe Ratio will also be scrutinized.
        - **Network Structure Stability:**The impact of different mitigation strategies on network topology (e.g., connection density, centrality distribution) will be analyzed. The key question is whether bias removal leads to significant changes in network connectivity, such as the elimination of seemingly plausible but anachronistic ”shortcut” connections.
        - **Known Relationship Recovery:**The alignment of networks from each group with external ground-truth relationship data (e.g., from FactSet Revere) will be compared. An ideal strategy should remove bias without unduly compromising the ability to capture genuine historical relationships.
        - **True Out-of-Sample Test:**Predictions will be tested on data that post-dates the model’s knowledge cutoff (e.g., using data from 2024 and beyond if the model was trained up to the end of 2023). In this scenario, look-ahead bias is theoretically absent, and the model’s performance can serve as a benchmark for its true capabilities, to be compared with its historical backtesting performance.
        - **Quantitative Metrics:**Statistical tests (e.g., t-tests, Wilcoxon rank-sum tests) will be used to determine if the differences in key performance indicators across the various strategy groups are statistically significant.

By implementing this rigorous, detailed, and verifiable process, this research aims to maximally strip away the potential look-ahead bias introduced by pre-trained models. This ensures that the constructed global financial business networks more authentically reflect historical economic linkages, thereby providing a solid and reliable foundation for subsequent financial analysis and empirical investigation. The methodological contributions from this part of the study are not only crucial for the present research but may also offer valuable guidance for other historical analysis studies in finance that leverage LLMs and embedding models.

3.8 Limitations and Future Work

While the research framework detailed herein is designed to advance the construction of global financial business networks by integrating state-of-the-art AI technologies, any ambitious research endeavor is necessarily accompanied by methodological caveats. This concluding section of the methodology aims to transparently identify these potential limitations, framing them not as impediments but as points of departure for future inquiry, thereby ensuring the rigor, clarity, and forward-looking nature of the research design.

3.8.1 Inherent Limitations of the Proposed Framework

The successful execution of this research is contingent upon navigating several inherent challenges that define the current frontiers of the field.

Data Availability, Quality, and Alignment The framework's efficacy is fundamentally predicated on the accessibility and quality of underlying data. Despite a comprehensive multi-source data acquisition strategy, the systematic procurement of high-quality, structured multimodal data—particularly historical images, charts, and executive audio/video for emerging markets and Small and Medium-sized Enterprises (SMEs)—remains a significant hurdle. The inherent sparsity, noise, and formatting heterogeneity of such data, along with residual errors from cross-lingual processing, can directly impact the training efficacy of the multimodal fusion framework and, consequently, the accuracy and completeness of the final network representations. The precise temporal alignment of historical multimodal data, in particular, poses a persistent and non-trivial challenge for constructing reliable dynamic networks.

Model Complexity, Interpretability, and Generalizability The research leverages sophisticated, yet inherently complex, embedding models and novel multimodal fusion architectures. A key challenge associated with these deep learning models is their ”black box” nature. While they may demonstrate superior performance in predictive tasks, providing a clear and intuitive explanation for why a specific connection forms between two firms, or quantifying the precise contribution of a single modal input (e.g., a product image or a vocal inflection), remains difficult. This lack of interpretability may constrain the direct applicability of the findings in high-stakes financial decision-making contexts, such as regulatory oversight or risk management, where transparency and trust are paramount. Furthermore, the generalizability of the trained models—their performance on unseen companies, industries, or market conditions—requires extensive and continuous validation.

Distinguishing Relationship Types and Causal Inference While the research design incorporates methods to differentiate between competitive, supply-chain, and other relationship types, any approach predicated on similarity metrics faces inherent limitations in reliably and precisely disentangling complex economic ties, especially in the absence of large-scale, high-quality labeled data. More fundamentally, the constructed networks primarily reflect **correlation**, not **causation**. Although extrinsic evaluation through downstream tasks can demonstrate the network's predictive value, inferring causal links directly from the observed network structure must be approached with extreme caution. The current framework cannot fully preclude the influence of confounding variables or address issues of endogeneity, which are necessary for robust causal claims.

Residual Look-ahead Bias Despite the implementation of a systematic, multi-pronged strategy to mitigate look-ahead bias, including the use of advanced NER and context-aware masking, the complete elimination of all potential future information latent within large pre-trained models is an exceedingly difficult, if not impossible, task. The potential for a subtle, residual bias to remain could still affect the absolute accuracy of historical network analysis, particularly in fine-grained time-series analyses or backtesting scenarios.

Computational Resource Constraints The entire methodological pipeline—from processing massive-scale global text and multimodal data to training and running advanced embedding and fusion models, and conducting large-scale network analysis and simulations—is computationally intensive. This requires substantial resources, including GPU clusters and large-scale storage, which not only imposes significant research costs but may also limit the scalability (e.g., coverage of more companies, higher-frequency updates) and real-time applicability of the framework.

3.8.2 Research Agenda for Future Work

These limitations illuminate a clear and compelling agenda for future research, positioning this study as a foundational step in a long-term research program.

Enhancing Data Foundations and Processing Capabilities Future work could explore the integration of emerging alternative data sources—such as satellite imagery for monitoring supply chain activity, patent databases for uncovering technological linkages, social media for analyzing public sentiment, or employee review platforms for gauging internal corporate dynamics—to supplement and validate the multimodal network. Research into more advanced data fusion and alignment techniques, particularly those leveraging self-supervised learning or generative models for data augmentation and cross-modal alignment in sparse, noisy historical contexts, would be highly valuable. Furthermore, developing more efficient, low-cost data processing pipelines through model compression and knowledge distillation could mitigate computational constraints.

Improving Model Interpretability and Robustness A critical avenue for future research is the systematic application of Explainable AI techniques (e.g., SHAP, LIME, attention visualization, Concept Activation Vectors) to the developed multimodal embedding and fusion models. This would help to uncover the key features and modal contributions that drive network connections, providing a more transparent basis for financial decision-making. Concurrently, research into graph neural network architectures designed with intrinsic interpretability for financial networks, or the development of more robust embedding and fusion methods resilient to data noise and adversarial attacks, would represent a significant advancement.

Deepening Network Analysis and Pursuing Causal Inference Moving beyond simple similarity networks, future studies could focus on developing network representation learning methods capable of explicitly modeling heterogeneous relationship types (e.g., Heterogeneous GNNs), thereby constructing multi-layered or heterogeneous information networks that concurrently represent competition, supply, and collaboration ties. A crucial next step is to integrate formal causal inference frameworks (e.g., leveraging libraries like DoWhy or EconML) with network analysis. This could involve using network structure to identify instrumental variables or to design quasi-experimental studies (e.g., exploiting network ”break points” or exogenous shocks) to more rigorously identify causal effects between economic behaviors. Furthermore, investigating continuous-time dynamic graph models could offer a more granular understanding of the micro-mechanisms driving network evolution, moving beyond discrete time snapshots.

Expanding Applications and Enabling Real-time Monitoring The enhanced networks developed in this study could be applied to a broader array of financial problems, including credit risk assessment, the explanation of asset pricing anomalies, and the development of early warning systems for systemic risk. Beyond finance, applications in industrial policy evaluation, regional economic development studies, and innovation diffusion pathway analysis are also conceivable. A long-term ambition would be to explore the construction of a near-real-time global business network monitoring system, integrating stream data processing and incremental learning techniques to provide dynamic support for risk management and trading decisions within financial institutions. Ultimately, the constructed networks could be integrated into larger-scale Financial Knowledge Graphs, enabling deep, multi-source information consolidation and reasoning.

Acknowledging these limitations does not diminish the value of the present study; rather, it underscores its significance as an exploratory and frontier-pushing endeavor. By systematically addressing existing gaps in the literature and prudently navigating methodological challenges, this research aims to provide a significantly improved analytical framework for understanding the complex interdependencies of the global economy. In doing so, it lays a solid foundation for a sustained and impactful long-term research agenda.

4 Expected Result

The central ambition of this research is to overcome the manifold limitations of existing studies in financial business network construction through systematic methodological innovation. The ultimate aim is to develop and validate an Enhanced Global Financial Business Network that demonstrates significant improvements in accuracy, comprehensiveness, dynamism, and demonstrable economic value. Following a detailed exposition of the research background, motivation, literature review, significance, and a meticulous research design and methodological framework, this section delineates the key outcomes and specific findings anticipated from the successful execution of the proposed research plan. These expected results not only directly address the five core objectives set forth in this study but also represent a concerted effort to break through current technological bottlenecks and knowledge gaps, thereby providing new empirical tools and theoretical perspectives for understanding the complex global economic and financial system. The subsequent discussion will elaborate on the anticipated specific outputs and their significance, structured around each research objective.

4.1 A High-Quality, Broad-Coverage Global Corporate Description Database

As the bedrock for all subsequent analyses in this study, the successful development and validation of an automated, scalable, and rigorously quality-controlled data processing pipeline is a primary anticipated outcome. The innovation of this pipeline lies in its synergistic combination of SOTA LLMs; refined prompt engineering techniques; the judicious (if activated) use of RAG; and multi-layered data validation mechanisms. This pipeline is expected to efficiently extract—or, where necessary, generate—highly standardized, information-rich, and precisely timestamped corporate business descriptions from a diverse array of heterogeneous global text sources, including annual reports, investor conference call transcripts, press releases, and regulatory filings. A key focus is the explicit extension of coverage to emerging markets and SMEs, which have been largely overlooked in prior research. Specifically, the anticipated outcomes are:

- **A Core Curated Dataset:**This research is expected to produce a core dataset comprising historical business descriptions for a vast number of global firms. Each description will be clearly time-stamped (e.g., corresponding to fiscal year-ends or report publication dates) and uniformly formatted to facilitate subsequent processing by embedding models. It is anticipated that this dataset will achieve a significant increase in global market capitalization coverage, striving to meet or exceed a 90% target. The coverage of emerging markets and SMEs is expected to far surpass that of existing public datasets or those described in the literature, thereby effectively addressing the systemic deficiencies in data breadth identified in the literature review.

- **Validation of the Superiority of the Data Processing Pipeline:**Through comparative evaluation against benchmark methods—such as relying solely on descriptions from commercial databases, simple extraction from raw text, or generation using earlier-generation LLMs—the proposed pipeline is expected to demonstrate marked advantages. These include superior information quality and semantic fidelity, as validated through manual expert reviews (targeting 80% alignment with financial analyst judgments) and semantic similarity comparisons with high-quality commercial database descriptions. The pipeline is also expected to exhibit enhanced standardization and consistency by effectively processing diverse source formats and languages into a comparable output. Crucially, through rigorous input filtering and targeted prompt constraints, the process is designed to ensure temporal consistency, meaning historical descriptions will be free of future information, laying a vital foundation for subsequent look-ahead bias mitigation. Finally, the automated workflow is anticipated to offer superior processing efficiency and scalability compared to manual or traditional NLP approaches.

- **A Replicable Methodological Framework:**A key deliverable will be a set of best practices for the responsible and effective use of LLMs in the large-scale, historical, structured information extraction and standardization of financial texts. This will include optimized prompt strategies, quality control protocols, and risk mitigation mechanisms (particularly for bias).

In essence, the successful realization of Objective 1 will not only furnish an unprecedentedly high-quality data foundation for the subsequent phases of this research but also, through the dataset and validated methodology it produces, make a direct and valuable contribution to the field of financial text analysis. The resulting assets have the potential to be a valuable resource for the broader academic and industry research communities (subject to data licensing agreements).

4.1.1 Optimal Embedding Model for Financial Network Construction

Addressing the current lacuna in the literature regarding the systematic, domain-adapted evaluation of the latest embedding models for financial network construction, the successful implementation of the framework for Objective 2 is expected to yield several critical outcomes:

- **A Validated, Finance-Oriented Embedding Model Evaluation Framework:**A primary deliverable will be a comprehensively documented and empirically validated multi-dimensional quantitative evaluation framework. This framework, specifically tailored for the task of financial business network construction, will transcend generic NLP metrics. It will encompass key dimensions such as semantic similarity capturing (e.g., correlation with human judgments, precision/AUC in recovering known relationships from sources like FactSet Revere), economic structure coherence (e.g., the plausibility of industry/country clusters as measured by AMI or Silhouette Score), and downstream financial task performance (e.g., predictive power in M&A forecasting or profitability in network-based trading strategy backtests). The framework itself will constitute a significant methodological contribution.

- **Rigorous Comparative Insights on Cutting-Edge Embedding Models:**By systematically comparing a suite of the latest commercial closed-source and open-source (e.g., SOTA Sentence Transformers models, locally deployed E5) embedding models, alongside necessary baselines (e.g., TF-IDF, Word2Vec), this research expects to produce statistically significant empirical findings on their relative strengths and weaknesses in capturing complex economic linkages. The results are anticipated to reveal how different model architectures, training data, or domain adaptations concretely affect their ability to capture specific financial semantics (e.g., competition vs. supply chain links) or to handle texts in various languages.

- **Identification of the Optimal Model or Model Ensemble for This Research:**Based on the rigorous comparative assessment, and balancing considerations of model performance (especially in known relationship recovery and downstream tasks), potential for economic interpretability, and computational efficiency and cost, a key outcome will be the clear identification and selection of a single model, or a synergistic ensemble of models, that achieves the best overall balance for the subsequent business network construction. This selection will be data-driven and justifiable, moving beyond the ad-hoc adoption of general-purpose models.

- **A Reusable Benchmark and Decision-Making Resource for the Field:**The detailed documentation of the evaluation framework, implementation procedures, comparative results, and the rationale for model selection is expected to form an important reference benchmark. This will provide a much-needed, empirically-backed decision-making basis and methodological reference for future researchers and practitioners in FinTech when selecting and applying embedding models for business network analysis or other related financial text tasks, thereby fostering greater standardization and comparability in the field.

In conclusion, the anticipated outcomes for Objective 2 extend beyond selecting the optimal technical tools for this study. By establishing a domain-adapted evaluation system and providing rigorous comparative evidence, this research aims to substantively fill a critical methodological gap in the existing literature and offer long-term reference value to the entire field.

4.1.2 An Effective Multimodal Fusion Framework

A primary anticipated result is the delivery of an operational multimodal joint embedding model. This deep learning model will be engineered to accept and process heterogeneous inputs from diverse sources, including text, images, audio (e.g., sentiment features from conference calls), and structured financial/ESG data. The architecture, with a particular emphasis on a contrastive learning approach, is expected to effectively map these disparate information streams into a unified, shared low-dimensional semantic space. The operational viability of the model will be demonstrated by its stable performance in processing large-scale global company datasets and its capacity to generate a fused embedding vector, v_{fused_t}, for each company at every time point. The successful convergence of the training process, such as the contrastive loss, will serve as initial evidence of the model's efficacy.

Crucially, this research expects to provide a quantifiable demonstration of the superiority of the multimodal fused representation. The core hypothesis is that the multimodal fused embedding vectors, v_{fused_t}, will be significantly superior to the optimal text-only embeddings, v_{text} (from Objective 2), in representing firms and their economic interconnections. This superiority will manifest as a more precise capture of economic relationships, verifiable through the evaluation framework:

Intrinsic Evaluation: It is expected that company similarities computed from v_{fused_t} will exhibit a higher correlation (Spearman's rank) with expert human judgments of business similarity. Furthermore, these fused embeddings are anticipated to achieve higher precision, recall, F1-scores, and AUC values in recovering known relationships from ground-truth benchmarks (e.g., FactSet Revere), particularly for relationships driven by non-textual cues (e.g., competition based on product visual similarity). Company clusters derived from v_{fused_t} are also expected to more clearly reflect economically intuitive industry and regional structures (evidenced by improved Silhouette Scores, AMI, etc.).

Extrinsic Evaluation: In downstream financial applications (see Objective 5), models utilizing v_{fused_t} as a core feature are expected to demonstrate statistically significant performance improvements over models using only v_{text}. This could include higher AUC in M&A prediction or better risk-adjusted returns in trading strategies based on lead-lag effects.

Beyond demonstrating overall superiority, the research anticipates revealing the key informational contributions of specific non-textual modalities. Through in-depth analysis of the fusion model, particularly by leveraging attention mechanisms or conducting targeted ablation studies, the specific contributions of different non-textual modalities (e.g., images, audio sentiment, structured data) to the accuracy of identifying particular types of economic relationships will be quantified. For instance, it is conjectured that product image embeddings will prove vital for identifying competitors in consumer goods industries, while vocal sentiment features from executive conference calls may offer unique value in assessing risk transmission or collaborative intent linked to management confidence. This will be measured by analyzing the model's internal attention weights or by quantifying the performance degradation observed in ablation studies when a specific modality is removed, thereby providing new empirical evidence on the precise role of various multimodal signals in financial markets.

Ultimately, the successful completion of this objective will provide an empirical solution to the literature gap in multimodal network construction. By not only proposing an innovative fusion framework but also rigorously demonstrating its efficacy and value-add over text-only approaches, this research will directly address the lack of methodology and evidence for systematically leveraging multimodal information to build large-scale financial business networks. This will furnish the FinTech field with a validated analytical tool for constructing enhanced multimodal global business networks, propelling research in this domain from a focus on single-company analysis toward a more sophisticated, network-based paradigm of relational analysis.

4.1.3 Validated and Effective Strategies for Look-ahead Bias Mitigation

This research expects to achieve clear and quantifiable outcomes in addressing the critical methodological challenge of look-ahead bias. Through systematic comparative experiments and rigorous evaluation, the proposed bias mitigation strategies, based on advanced NER and context-aware masking, are expected to be validated as effective.

A key anticipated outcome is the quantitative demonstration of significant bias mitigation. By comparing historical networks constructed using different processing strategies (no bias treatment, simple masking, and context-aware masking), statistically significant differences are expected in downstream applications requiring historical backtesting, such as lead-lag trading strategies. Specifically, a significant and justifiable reduction in the potentially inflated predictive power or excess returns (alpha) of the baseline (untreated) network is anticipated when the optimized bias control strategy is applied. This reduction would not signify a failure of the strategy but, on the contrary, would provide strong evidence of the successful removal of spurious signals originating from ”future information,” thereby restoring the historical authenticity of the results.

The evaluation is also expected to elucidate the trade-off between bias removal and information preservation, leading to best practices. By comparing how different strategies affect the network's ability to match known historical relationships (e.g., completed M&A deals, confirmed supply chain links), it is anticipated that the proposed context-aware masking will prove superior to simple deletion or anonymization, as it should better preserve the semantic information and network structure reflecting genuine economic ties while effectively removing bias. Based on this quantitative evidence, a set of best-practice recommendations or operational guidelines will be proposed for achieving an optimal balance between bias removal and information fidelity, offering a valuable reference for other researchers responsibly applying large pre-trained models to historical financial analysis.

Ultimately, through the application of these validated bias control strategies, the reliability and credibility of the historical network analysis will be substantially enhanced. This will directly improve the trustworthiness of all subsequent empirical analyses based on the historical networks—including dynamic evolution studies, event impact analyses, and strategy backtesting—forming a critical cornerstone for the validity of the study's overall conclusions.

4.1.4 Empirical Value of the Enhanced Network

This objective serves as the core test of the practical utility of the research's final output: the enhanced global financial business networks (G_{text} and G_{multi}). Through rigorous empirical examination across several key financial application scenarios, the constructed networks are expected not only to statistically and significantly outperform traditional benchmark methods (e.g., those based on industry classification, stock return correlations, or earlier text-based approaches) but also to reveal new economic patterns, provide deeper insights, and potentially generate actionable strategic implications.

In the analysis of global stock return lead-lag effects, the uncovering of finer and stronger cross-firm information transmission mechanisms is anticipated. Compared to benchmark networks, the linkage strengths (embedding similarities) and centrality measures derived from G_{text} and particularly G_{multi} are expected to more significantly and robustly predict future stock return lead-lag relationships. It is anticipated that novel, and perhaps faster, information transmission pathways driven by specific multimodal signals—such as shifts in managerial sentiment in conference calls or the popularity of product launch videos—will be identified. Quantitatively, investment portfolio strategies based on these networks (e.g., pairs trading using connection strength, or stock selection based on centrality) are expected to yield statistically significant and economically meaningful risk-adjusted excess returns (alpha), which should be demonstrably higher than those from similar strategies based on benchmark networks. This would validate the value of higher-precision semantic understanding and multimodal information fusion in capturing financial market information flows, moving beyond traditional methods that rely solely on price signals or coarse textual similarities.

A significant improvement in the early prediction accuracy of corporate M&A is expected. By incorporating network features extracted from G_{text} and G_{multi} (e.g., firm-pair embedding similarity, network distance, common neighbor features, changes in node centrality) into standard M&A prediction models, a significant enhancement in the predictive accuracy (measured by AUC and F1-score) for future, especially cross-border, M&A events is anticipated. Multimodal features, such as product visual similarity or technological patent text associations, are expected to provide unique early warning signals. Quantitatively, the network-derived features are expected to exhibit significant marginal predictive power after controlling for traditional predictors like firm financials, size, and industry. This would demonstrate that a granular characterization of inter-firm business similarity and strategic relatedness is crucial for understanding M&A motivations, and that multimodal information offers incremental value overlooked by traditional text analysis.

A more precise quantification and tracking of the propagation pathways and impact of ESG spillovers is anticipated. When using the spatial weight matrices derived from G_{text} and G_{multi} in spatial econometric models (e.g., SAR, SDM), statistically significant and economically meaningful spillover effects of ESG performance (e.g., score changes, positive/negative events) across the business network are expected to be found. The research anticipates identifying differentiated propagation pathways and intensities of ESG impacts through different types of connections (e.g., supply chains vs. competitors) or via different modal signals (e.g., environment-related visual evidence vs. governance-related textual disclosures). Quantitatively, the estimated spatial autoregressive coefficients (ρ or λ) are expected to be significantly non-zero, with their signs and magnitudes providing a plausible explanation of ESG contagion mechanisms. This would offer more granular and direct empirical evidence for the ”impact” of ESG investing, moving beyond traditional firm- or industry-level analysis.

A more operationally relevant assessment of global supply chain resilience and risk identification is expected. The more granular global supply chain map constructed from G_{text} and

G_{multi}—incorporating features like product images and geographical information—is expected to reveal more complex and vulnerable cascading failure pathways under network perturbation simulations (e.g., removing key nodes/edges) than those identifiable from traditional supply chain data (e.g., Bloomberg SPLC) or industry classifications. The identification of previously underappreciated, ”hidden” keystone supplier or customer nodes is a key anticipated finding. Quantitatively, the simulated network robustness metrics are expected to align more closely with real-world observations (e.g., actual supply chain disruptions following specific events), leading to more accurate quantitative measures of supply chain risk exposure that can inform corporate risk management and diversification strategies. This would showcase the power of the proposed network in characterizing complex economic systems, especially in areas like supply chains where standardized data is scarce.

By achieving these anticipated outcomes across four highly relevant and academically valuable application scenarios, this research will provide strong evidence that the constructed enhanced global financial business network not only represents a technological breakthrough but, more importantly, offers substantial incremental value in understanding market dynamics, predicting key events, and assessing emerging risks, thereby validating the efficacy and importance of the entire research framework.

In summation, should the anticipated results of this research be successfully realized, they will not only address each of the five core research objectives but, more importantly, will synergistically constitute a systematic improvement to the existing paradigm of global financial business network construction and analysis. By developing and validating innovative data processing pipelines, establishing a finance-oriented benchmark for embedding model evaluation, constructing a cutting-edge multimodal fusion framework, implementing more effective bias control strategies, and ultimately demonstrating superior empirical value across multiple key financial applications, this research is expected to deliver an enhanced global business network and its associated methodology that significantly surpasses existing methods—whether traditional industry classifications, market-data-based approaches, or the latest text-embedding methods—in terms of accuracy, coverage, informational depth, temporal reliability, and economic relevance. The realization of these outcomes will directly fill the critical gaps identified in the literature review, namely the challenges in handling data heterogeneity and coverage limitations, the lack of standardized model evaluation, the insufficient integration of multimodal information, the difficulty in controlling look-ahead bias, and the inadequate depth of application. Ultimately, the anticipated outcomes will not only serve as a powerful validation of the proposed research design’s effectiveness and innovation but will also provide the academic community with a more potent analytical toolkit for understanding complex global economic interdependencies and furnish financial practitioners with more reliable and deeper insights for investment decisions, risk management, and strategic formulation, thereby taking a solid step forward at the research frontier of the field.

5 Feasibility Analysis

After having systematically articulated the background, motivation, literature review, significance, and expected outcomes of this research, alongside a detailed research design and methodological framework, this chapter shifts its focus to a pragmatic assessment of the entire research plan’s Practical Feasibility. The preceding sections have delineated an ambitious research blueprint, situated at the forefront of the intersection between financial technology and artificial intelligence. This plan involves processing large-scale, multimodal, cross-border datasets; applying SOTA embedding and LLMs; developing innovative multimodal fusion architectures; and tackling complex methodological challenges such as the systematic control of look-ahead bias. Given the technical depth and breadth of this research agenda, a comprehensive, objective, and transparent analysis of its feasibility is not only necessary but also a critical step in ensuring the successful execution of the study and the ultimate achievement of its intended objectives.

Therefore, this feasibility analysis aims to demonstrate that, despite its significant innovation and inherent challenges, the proposed research plan is highly viable and can be effectively executed within a standard doctoral program timeline. This conclusion is based on a careful consideration of available resources, technological maturity, the research team's capabilities, and potential risks. The assessment will be detailed across several key dimensions:

- **Data Feasibility:**Covering the accessibility of multi-source heterogeneous data, the complexity of its processing, and associated costs.
- **Methodological and Technical Feasibility:**Evaluating the implementability and maturity of core AI models, network construction techniques, and analytical tools.
- **Computational Resource Feasibility:**Examining the availability of necessary high-performance computing and storage resources.
- **Expertise and Support Feasibility:**Considering the researcher's foundational skills, the expert guidance of the supervisory team, and the supportive institutional environment.
- **Project Timeline Feasibility:**Presenting a realistic and phased execution schedule.
- **Risk Assessment and Mitigation Strategies:**Identifying potential obstacles and formulating contingency plans.

Through an in-depth analysis of these dimensions, this chapter aims to provide robust support for the soundness and credibility of the entire research proposal, demonstrating that it is not merely an attractive academic concept but a tangible research practice with a clear implementation pathway.

5.1 Data Feasibility

The core of this research plan is the construction of an unprecedented global financial business network that integrates multimodal information, a task that places exceptionally high demands on the breadth, depth, quality, and timeliness of the data. The challenge of acquiring and processing a dataset that spans major global markets (including emerging markets and SMEs), covers over two decades, and encompasses multiple modalities—text, images, audio, and structured data—is fully acknowledged. Nevertheless, based on a prudent assessment of available data resources, advanced processing technologies, and associated costs, the data requirements for this study are deemed highly feasible.

- **Availability and Accessibility of Data Sources**

  The data foundation of this research relies on a diversified acquisition strategy, the feasibility of which is predicated on the effective utilization of multiple data access channels.

  - **Core Textual and Structured Data:**The acquisition of core data will be secured through several avenues. Institutional subscriptions to premier financial and business databases (such as LSEG Refinitiv Eikon/Workspace, Bloomberg Terminal, FactSet, S&P Capital IQ, Compustat, CRSP, I/B/E/S) will be leveraged extensively. These databases provide access to standardized financial statements, market data, analyst forecasts, ESG scores, some supply chain information, and curated news and basic company information for major global markets, forming the cornerstone for high-quality structured data and some standardized text. For US market data, the SEC EDGAR database offers a free and comprehensive public access interface, which will be utilized through automated scripts leveraging established Python libraries to download key filings like 10-K and 10-Q reports. For other major international markets, public data interfaces or websites from national regulatory bodies (e.g., ESMA in Europe, CSRC in China) will be explored. High-quality transcripts of investor conference calls will be sourced from commercial providers, often available as modules within the subscribed databases.
  - **Multilingual and Global Coverage:**By combining the use of the mentioned global commercial databases with advanced multilingual processing techniques (as detailed in Section 2.2, it is believed that language barriers can be effectively overcome to achieve broad geographical coverage, including major emerging markets. This directly addresses the limitations of regional coverage noted in the literature.
  - **Multimodal Data Acquisition:**It is recognized that the systematic acquisition of high-quality, structured, and precisely timestamped historical multimodal data (especially images and audio) is a more challenging endeavor. The strategy here will be pragmatic and phased. The initial priority will be to extract publicly available images and potential video links embedded within company annual reports, official websites, and product brochures, using PDF parsing and web scraping technologies. This constitutes a technically mature and feasible baseline source for multimodal data. Furthermore, some commercial providers are beginning to offer structured audio files for conference calls or related metadata, and these resources will be actively utilized. For more exploratory avenues, such as satellite imagery, publicly available datasets will be leveraged, or targeted commercial data procurement will be considered if budget allows. It is anticipated that the coverage and historical depth of multimodal data may not match that of textual data. However, this does not preclude an exploration of its incremental value. The initial stages of the research will focus on recent data and markets with higher-quality information disclosure, and the potential impact of uneven data coverage will be explicitly considered in the analysis.

- **Feasibility of Data Processing**

  Transforming raw, heterogeneous, and massive datasets into a clean, structured format suitable for advanced AI model input is a technical cornerstone of this research. The complex preprocessing pipeline outlined in the Methodology Framework is considered entirely feasible from a technical standpoint. The entire workflow will be built upon a mature and powerful open-source ecosystem centered on Python. Standard libraries and frameworks, including but not limited to Pandas and NumPy for data manipulation, Requests and Scrapy for data acquisition, PyMuPDF and pdfminer.six for PDF parsing, spaCy and Hugging Face Transformers for NLP tasks (tokenization, NER, multilingual processing), OpenCV for image processing, and Librosa for audio analysis, will be utilized. The research team possesses, or will rapidly acquire, the necessary proficiency in these tools.

Key complex processing steps have clear and manageable solutions. Cross-lingual processing can be effectively handled using high-quality commercial translation APIs or powerful multilingual models from the Hugging Face library. The core challenge of entity resolution and linking across disparate data sources will be addressed primarily through exact matching based on standard identifiers (ISIN, CUSIP, LEI), supplemented by fuzzy string matching algorithms, external reference databases (like OpenCorporates), and potentially machine learning methods to handle name variations and missing identifiers. Strict data annotation protocols will be implemented to assign both ”event time” and ”as-of time” stamps to all data points, a process that is fundamental for the subsequent effective handling of look-ahead bias.

- **Manageability of Data Costs**

  The execution of such a large-scale, data-intensive study inevitably involves cost considerations, primarily related to commercial database subscriptions and potential API usage fees. Access to core commercial databases is largely covered by existing institutional-level subscriptions provided by college or department, which controls the majority of the basic data costs within the conventional research funding framework. For API services that may incur additional fees (e.g., LLM generation, high-quality translation, specific embedding model APIs), a prudent management strategy will be adopted. This includes pre-allocating a dedicated budget for API calls, optimizing for cost-effectiveness by prioritizing lower-cost API options or model versions for non-critical tasks, and leveraging powerful open-source alternatives to reduce reliance on paid APIs. API usage will be strictly controlled through optimized code (e.g., batch requests), caching mechanisms, and the use of sampled data during development and testing. Furthermore, academic research credits or educational discounts from cloud service providers will be actively pursued.

In summary, although the data requirements for this research are demanding, a strategic combination of leveraging institutional resources, employing mature open-source technologies, implementing a rigorous data processing pipeline, and effectively managing costs ensures that the necessary data acquisition and processing are practically feasible. This provides a solid data foundation for achieving the subsequent research objectives.

5.2 Methodological and Technical Feasibility

The core of this research plan lies in the application and integration of a suite of cutting-edge AI technologies—including LLMs, advanced embedding representations, and multimodal fusion—to construct and analyze global financial business networks. A prudent assessment of the feasibility of the key technologies underpinning these methodologies leads to the conclusion that, despite inherent challenges, they are both implementable and operationally viable within the current technological landscape.

- **Implementability of Core AI Models**

  The application of LLMs for the high-quality extraction or generation of company descriptions is highly feasible. Access to these state-of-the-art models is readily available through mature and well-documented commercial APIs from providers, which obviates the need for self-training or maintaining extensive infrastructure. Concurrently, high-performance open-source LLMs (e.g., Llama, Mistral series) provide options for local deployment, contingent on computational resource availability. The central research challenge in this domain is not one of basic access but of refined prompt engineering and rigorous quality control of the generated content. These are advanced skills that, while requiring iterative optimization, can be systematically developed and perfected through experimentation within the scope of a doctoral research project.

Similarly, the implementation of advanced embedding models is highly feasible. Commercial models can be accessed via their respective APIs to generate embedding vectors directly. A multitude of top-tier open-source embedding models are conveniently accessible through widely adopted Python libraries like Hugging Face ‘transformers‘ and ‘sentence-transformers‘, allowing for efficient inference on local GPU environments or cloud platforms. This has become a standard workflow in contemporary NLP and machine learning research.

The development of the multimodal fusion framework, while representing the most technically challenging and innovative aspect of this study, is also feasible. The proposed fusion strategies are grounded in active and established research areas, with a wealth of public literature, open-source codebases (e.g., PyTorch, TensorFlow, and their ecosystems), and mature theoretical foundations to draw upon. The core task is not to invent entirely new foundational model architectures from scratch but rather to intelligently design, adapt, and combine existing technological modules—such as pre-trained text, image, and speech encoders, attention mechanisms, and contrastive loss functions—to suit the specific characteristics of financial data and the objective of business network construction. This endeavor is inherently an exploratory yet well-defined task appropriate for doctoral research, with ample time allocated for architectural design, experimental iteration, and performance optimization.

Finally, the implementation of the proposed strategies for controlling look-ahead bias (utilizing advanced NER for entity recognition and context-aware masking) is technically straightforward. A variety of mature NER toolkits (e.g., spaCy, Flair, Microsoft’s Presidio, and numerous pre-trained NER models on Hugging Face) are available for use. Context-aware masking is essentially a text processing task based on NER output, which can be implemented with standard Python scripts. The innovation here lies in the design of sophisticated masking strategies and the systematic evaluation of their effectiveness, rather than in the development of the underlying tools.

- **Availability of Network Construction and Analysis Tools**

  The process of constructing graph structures from embedding vector similarities and conducting subsequent topological analysis will rely entirely on mature and widely used open-source Python libraries. For instance, NetworkX offers comprehensive tools for graph creation, manipulation, and basic algorithm implementation, while ‘igraph‘ and ‘graph-tool‘ provide higher performance for handling very large-scale networks. Using these libraries to compute degree distributions, clustering coefficients, centrality measures, and perform community detection (e.g., with built-in Louvain algorithms) is standard practice.

  The econometric and machine learning models required for the downstream applications in the evaluation framework (e.g., M&A prediction, ESG spillover analysis, backtesting of return prediction strategies) are also well-supported by powerful open-source libraries. The `statsmodels` library provides classic regression and time-series analysis tools; `PySAL` is dedicated to spatial econometric models (like SAR and SDM); `scikit-learn` offers a wide array of machine learning algorithms including Logistic Regression and SVM; and `XGBoost` and `LightGBM` provide efficient implementations of gradient boosting trees. The implementation of the empirical analysis models designed in this study using these libraries is not only feasible but also standard practice in quantitative finance and economics research.

In summary, the core methodologies and technical tools required for this research plan, from advanced AI models to standard network and data analysis libraries, all have reliable implementation pathways. The availability of commercial APIs and mature open-source libraries significantly lowers the technical barrier to entry, allowing the research focus to be on the innovative application of methods, their adaptation and optimization for financial contexts, and rigorous empirical evaluation, rather than on the redundant development of foundational tools. Therefore, from a methodological and technical standpoint, this research plan is highly feasible.

5.3 Computational Resource Feasibility

The execution of this research plan, particularly the processing of massive global datasets, the operation of advanced LLMs and embedding models, the training and application of the multimodal fusion framework, and large-scale network analysis and simulation, places significant demands on computational resources. These requirements have been carefully assessed, and corresponding resource acquisition and management strategies are in place, ensuring the high feasibility of the study from a computational perspective.

- **Assessment of Computational Requirements**

  The key computationally intensive tasks in this study include:

  - **Large-Scale Data Processing and Storage:**The processing of an estimated multi-terabyte dataset of raw multi-source, multimodal data requires powerful data handling capabilities and vast storage space.
  - **Model Inference:**Applying the latest LLMs for generating standardized business descriptions will require considerable computational power (especially for local deployment) or a sufficient budget for API calls. Generating high-quality embedding vectors for tens of thousands of global companies across multiple time points will necessitate efficient GPU-accelerated inference, particularly for feature extraction from multimodal data using models like CLIP, ViT, and Wav2Vec2.
  - **Training and Fine-tuning:**The training of the multimodal fusion framework, especially when exploring architectures based on contrastive learning or complex attention mechanisms, will be one of the most computationally demanding aspects, requiring high-end GPUs with large memory (e.g., 40GB or more, such as NVIDIA A100 or H100) for extended training periods.
  - **Large-Scale Network Analysis and Simulation:**Computing embedding similarities for millions or even billions of firm-pairs, constructing and storing large networks, and running graph algorithms will be resource-intensive. The estimation and simulation of empirical models for downstream applications may also require multi-core CPUs or GPU acceleration.

- **Assessment of Computational Resources**

  To meet these high computational demands, several avenues of support are anticipated:

  - **Institutional High-Performance Computing Cluster:**Access to the university's HPC cluster is expected. These GPU resources are crucial for model training (especially for the multimodal fusion) and large-scale embedding inference. The HPC also provides large-scale parallel processing capabilities and high-speed network storage, which are more than adequate for the data processing and analysis needs of this study.
  - **Cloud Computing Resources:**Cloud service credits (e.g., from AWS, Google Cloud, Azure) can serve as a valuable supplement to the HPC, especially when specific instance types (e.g., very large memory instances), elastic scaling, or particular cloud-native AI services (e.g., Google Vertex AI's Gecko embedding API) are required.
  - **Lab/Workstation Resources**The applicant's personal possession of a desktop workstation with a high-performance GPU and a laptop suitable for basic development and testing will facilitate day-to-day research efficiency, including code development, small-scale data testing, and model prototyping.

- **Software Environment and Technology Stack**

  The core software required for this research consists of mature, widely used open-source tools or software accessible through institutional licenses, ensuring high feasibility of the technology stack. The primary programming language will be Python, supported by its rich scientific computing ecosystem.

- **Resource Management and Optimization Strategy**

  To ensure the efficient use of valuable computational resources within budget and time constraints, several strategies will be employed:

  - **Code Optimization:**Writing efficient, vectorized code and leveraging the underlying optimizations of libraries like NumPy and Pandas. For computationally intensive parts, GPU-accelerated libraries like CuPy or Faiss will be considered.
  - **Batch Processing and Parallelization:**Data processing, model inference, and some network analysis tasks will be accelerated through batch processing and parallel computation.
  - **Model Selection and Efficiency Trade-offs:**During the embedding model selection phase, computational efficiency and cost will be key considerations. Where performance is comparable, more lightweight or faster-inference models will be prioritized.
  - **Resource Monitoring and Scheduling:**Computational resource consumption will be closely monitored, and job priorities and resource requests will be managed according to HPC or cloud platform scheduling policies to avoid waste.
  - **Prioritization and Sampling:**During the exploratory phase or when facing extremely large computational loads, experiments and model tuning may first be conducted on representative data subsets before scaling up to the full dataset.

Although the computational requirements for this research are high, through clear access to resources (institutional HPC, cloud computing, local workstations) and systematic management and optimization strategies, the computational challenges are considered manageable and feasible. Sufficient computational power is a key guarantee for the successful processing of large-scale complex data, the application of advanced AI models, and the ultimate achievement of the research objectives.

5.4 Expertise and Support Feasibility

The successful execution of this research project is contingent not only upon a sound research design and adequate resources but also, critically, on the expertise and skills of the researcher and the available support system. This section argues for the high feasibility of the study from the perspectives of the applicant's capabilities, supervisory guidance, and institutional support.

5.4.1 Researcher's Background and Skills

A solid academic foundation combined with continuous research practice provides the applicant with the key skills and knowledge base necessary to undertake this study with confidence.

- **Interdisciplinary Knowledge Base:**The applicant holds dual master's degrees in Finance (from The Chinese University of Hong Kong, Shenzhen) and Computer Science (from The University of Hong Kong), enabling a fluent integration of financial theory with quantitative analytical methods. This background facilitates a deep understanding of the complexities of financial markets and corporate behavior, as well as the effective application of advanced computational techniques to solve problems in the financial domain.

- **Experience in NLP Research:**The applicant possesses a solid understanding of fundamental NLP theory and techniques, including text preprocessing, word embeddings, Transformer models, and tasks such as text classification and clustering. Practical experience in applying machine learning models to real-world problems enables the appropriate selection of models, effective feature design, and rigorous model evaluation and optimization. Specific expertise in LLMs includes:
  - Familiarity with the fundamental principles and evolving trends of LLMs, including key techniques like Prompt Engineering and Few-shot Learning.
  - Practical experience using LLM APIs for text generation, semantic understanding, and information extraction.
  - A demonstrated track record of relevant research, including a forthcoming publication at ACL 2025 on the vulnerability of LLMs to ”poisoned pills” data. Ongoing work as a Research Assistant (RA) at the School of Management and Economics / School of Data Science at CUHK-Shenzhen has provided continuous engagement in the application of LLMs in finance, accumulating valuable research and project experience.

- **Knowledge and Experience in the Financial Domain:**A strong theoretical foundation in financial markets, corporate finance, investment analysis, and risk management is complemented by familiarity with financial data sources and processing methods. The RA position has facilitated frequent interaction with faculty and researchers from other institutions, ensuring an up-to-date understanding of the latest research trends and frontier questions in finance.

- **Rapid Learning and Adaptability:**The applicant possesses a proven ability for continuous learning and rapid acquisition of new knowledge, essential for staying at the forefront of a fast-evolving technological field. There is a strong confidence in the ability to quickly master any as-yet-unfamiliar techniques or methods required for this research (such as advanced multimodal data fusion or sophisticated graph neural networks) through coursework, literature review, and hands-on exploration.

5.4.2 Planned Training

Despite a solid foundation of knowledge and practical experience, a plan for continuous skill development during the doctoral program is in place to further enhance the capabilities required for this research:

- **Relevant Coursework:**Enrollment in advanced courses such as Advanced Econometrics, Generative Models, Frontiers in FinTech, and Behavioral Finance to systematically deepen theoretical knowledge and methodological skills.
- **Participation in Workshops and Academic Conferences:**Attendance at relevant academic conferences and workshops to stay abreast of the latest research advancements and technological trends.
- **Online Learning and Practice:**Leveraging online learning platforms like Coursera, edX, and Udacity for targeted skill acquisition, and actively participating in open-source projects and relevant competitions to enhance practical abilities.

This analysis clearly demonstrates that the applicant not only possesses a strong knowledge base, outstanding research capabilities, and a fervent research interest but will also benefit from comprehensive support from the supervisor, research team, and institution. This combination of individual capability and external support provides a solid guarantee for the successful execution of this research.

5.5 Project Timeline Feasibility

This research is designed to be completed within a 12 to 18-month timeframe, necessitating a highly focused and efficient iterative research strategy. The following timeline, based on a careful assessment of the time and resources required for each task, highlights key milestones and accounts for potential time-related risks.

- **Phase 1: Data Acquisition and Preparation (Months 1-3)**
  - **Objective & Milestones**
    - Confirm and test access to all core data sources (commercial APIs, public databases, web crawlers).
    - Complete the download of raw data (annual reports, conference call transcripts, etc.) for at least 50% of the target companies by market capitalization.

    - ∗ Establish a preliminary data cleaning and preprocessing pipeline.
    - ∗ Initiate small-scale exploratory data analysis

  - **Key Deliverables**
    - ∗ Operational data scraping and parsing scripts.
    - ∗ A preliminary data quality assessment report.
    - ∗ A detailed inventory of data sources and access methods.
    - ∗ Clearly defined data quality evaluation metrics and procedures.

- **Phase 2: Text Embedding Model Evaluation and Selection (Months 3-6)**
  - **Objective & Milestones**
    - ∗ Complete the initial screening of candidate embedding models (both API-based and open-source).
    - ∗ Construct and test all components of the evaluation framework.
    - ∗ Run the evaluation framework on a preliminary dataset to obtain performance metrics for each model.
    - ∗ Finalize the selection of the optimal text embedding model or model ensemble.
  - **Key Deliverables**
    - ∗ An operational embedding model evaluation framework.
    - ∗ A comprehensive model evaluation report detailing the performance of each model on all metrics.
    - ∗ A clear and well-supported rationale for the selection of the optimal embedding model.

- **Phase 3: Multimodal Data Fusion and Network Construction (Months 3-10)**
  - **Objective & Milestones**
    - ∗ Complete the acquisition and preprocessing pipeline for multimodal data (images, audio, structured data).
    - ∗ Design and implement the multimodal fusion framework.
    - ∗ Complete the construction of the initial global business network.
    - ∗ Conduct small-scale network structure analysis and visualization.
  - **Key Deliverables**
    - ∗ Operational code for processing multimodal data.
    - ∗ Code implementing the multimodal fusion framework.
    - ∗ An initial global business network dataset (node and edge lists).
    - ∗ A report on network structure analysis and visualization.

- **Phase 4: Look-ahead Bias Control and Network Validation (Months 9-12)**
  - **Objective & Milestones**
    - ∗ Implement the look-ahead bias control process (NER + Masking).
    - ∗ Quantitatively evaluate the effectiveness of the bias control strategies.

    - ∗ Validate the economic value of the network in selected downstream financial tasks.
    - ∗ Finalize the main research findings.
  - **Key Deliverables**
    - ∗ An implementation plan and evaluation report for the bias control process.
    - ∗ An empirical results analysis report for the downstream financial tasks.
    - ∗ Detailed research findings and conclusions.

- **Phase 5: Dissertation Writing and Refinement (Months 12-15)**
  - **Objective & Milestones**
    - ∗ Draft the research dissertation.
    - ∗ Conduct internal reviews and revisions.
    - ∗ Submit the dissertation and prepare for defense (optional).
  - **Key Deliverables**
    - ∗ A first draft of the research dissertation.
    - ∗ The final, submitted research dissertation.
    - ∗ A public presentation of the research.

This is a highly compressed and ambitious timeline that will require stringent time management, effective collaboration, and adaptability. Close communication with the supervisor will be maintained through regular progress reports to allow for timely adjustments. Should any task require more time than allocated, priority will be given to research quality, with trade-offs made among other tasks to ensure the core research objectives are met within the 18-month timeframe. Through diligent planning and persistent effort, there is strong confidence in the ability to complete this challenging yet highly valuable research project on time and to a high standard of quality.

6 Conclusion

In an increasingly interwoven global economy, where the interdependencies between firms have grown profoundly complex, the ability to accurately and dynamically characterize the global financial business network has become a foundational prerequisite for understanding market structures, assessing systemic risk, and formulating effective investment and regulatory strategies. However, as the preceding review of the literature has demonstrated, prevailing methodologies—whether reliant on traditional static industry classifications, noise-prone market data, or early-stage textual analysis—exhibit severe and growing limitations in their capacity to precisely delineate the intricate, dynamic, and transnational nature of real economic linkages. These approaches fall short of the demands for depth, breadth, and timeliness required by modern financial analysis.

Positioned at the intersection of financial technology and artificial intelligence, this research confronts these challenges by proposing and developing an innovative methodological framework grounded in the latest advancements in AI, including LLMs, sophisticated embedding representations, and multimodal information fusion. The central objective is to construct and validate a global financial business network that achieves an unprecedented level of enhancement across accuracy, coverage, informational dimensionality, dynamism, and reliability—with a particular emphasis on the rigorous control of look-ahead bias. The successful realization of this objective is intended to provide a powerful analytical apparatus for profoundly understanding and effectively navigating the complex phenomena of the global economic and financial system.

This research is structured around five core, interconnected objectives. It commences with the development of a high-quality, broad-coverage, and temporally consistent global database of corporate business descriptions, designed to rectify the data insufficiencies and heterogeneity inherent in prior work. Upon this foundation, a finance-oriented evaluation benchmark will be established to scientifically select the optimal embedding model or model ensemble, thereby ensuring the semantic accuracy and financial relevance of the network's construction. The study will then proceed to design and implement an innovative multimodal fusion framework, integrating textual, visual, auditory, and structured data to significantly enrich the representation of inter-firm relationships. This entire process will be underpinned by the development and validation of effective strategies to control for look-ahead bias, safeguarding the veracity and reliability of historical network analysis. Finally, the demonstrable economic value of the constructed network will be ascertained through rigorous empirical testing across multiple key financial application scenarios, including stock return prediction, M&A event forecasting, ESG spillover analysis, and supply chain resilience assessment.

The anticipated outcomes of this research are poised not only to fill several critical lacunae in the current literature on global financial business network construction but also to catalyze significant progress in both the theory and practice of financial technology. Theoretically, this study will contribute a systematic framework for embedding model evaluation and a novel methodology for multimodal fusion, fostering a deeper integration of financial text analysis and network science. Practically, the research outputs will furnish investors, corporate managers, and regulatory bodies with more precise decision-support tools, enabling them to more effectively identify economic risks, capitalize on market opportunities, and enhance their risk management capabilities. On a societal level, the research supports the scientific assessment of ESG investments and promotes the responsible application of artificial intelligence, contributing to the sustainable development of financial markets.

Grounded in a solid theoretical foundation, an innovative technical roadmap, a clear implementation pathway, and with access to sufficient resources, this research is well-positioned to succeed. It is anticipated that this study will substantially elevate the quality of global financial business network construction and deepen its application, contributing a critical toolkit for understanding and addressing the ever-more complex challenges of the global economy and propelling the field of financial technology to new heights.

References

[1] Juan Alcácer, John Cantwell, and Lucia Piscitello. Internationalization in the information age: A new era for places, firms, and international business networks? *Journal of International Business Studies*, 47(5):499–512, 2016.

[2] Dogu Araci. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models, 2019.

[3] Nicholas Barberis, Andrei Shleifer, and Jeffrey Wurgler. Comovement. *Journal of Financial Economics*, 75(2):283–317, 2005.

[4] Brian Boyer, Todd Mitton, and Keith Vorkink. Expected Idiosyncratic Skewness. *Review of Financial Studies*, 23(1):169–202, 2010.

[5] Christian Breitung and Sebastian Müller. Global Business Networks. *Journal of Financial Economics*, 166:104007, 2025.

[6] Paul Brockman, Xu Li, and S. McKay Price. Conference Call Tone and Stock Returns: Evidence from the Stock Exchange of Hong Kong. *Asia-Pacific Journal of Financial Studies*, 46(5):667–685, 2017.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[8] Sean Shun Cao, Wei Jiang, Lijun (Gillian) Lei, and Qing (Clara) Zhou. Applied AI for finance and accounting: Alternative data and opportunities. *Pacific-Basin Finance Journal*, 84:102307, 2024.

[9] Xiangyu Chang, Lili Dai, Lingbing Feng, Jianlei Han, Jing Shi, and Bohui Zhang. A good sketch is better than a long speech: Evaluate delinquency risk through real-time video analysis. *Review of Finance*, 29(2):467–500, 2025.

[10] Lauren Cohen and Andrea Frazzini. Economic Links and Predictable Returns. *The Journal of Finance*, 63(4):1977–2011, 2008.

[11] Noah Cohen Kalafut, Xiang Huang, and Daifeng Wang. Joint variational autoencoders for multimodal imputation and embedding. *Nature Machine Intelligence*, 5(6):631–642, 2023.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[13] Assaf Eisdorfer, Kenneth Froot, Gideon Ozik, and Ronnie Sadka. Competition Links and Stock Returns. *The Review of Financial Studies*, 35(9):4300–4340, 2022.

[14] Matthew Elliott, Benjamin Golub, and Matthew O. Jackson. Financial Networks and Contagion. *American Economic Review*, 104(10):3115–3153, 2014.

[15] Eugene F. Fama and Kenneth R. French. Industry costs of equity. *Journal of Financial Economics*, 43(2):153–193, 1997.

[16] Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91:424–444, 2023.

[17] Evan Gatev, William N. Goetzmann, and K. Geert Rouwenhorst. Pairs Trading: Performance of a Relative-Value Arbitrage Rule. *The Review of Financial Studies*, 19(3):797–827, 2006.

[18] Paul Glasserman and Caden Lin. Assessing Look-Ahead Bias in Stock Return Predictions Generated By GPT Sentiment Analysis, 2023.

[19] Gerard Hoberg and Gordon Phillips. Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis. *The Review of Financial Studies*, 23(10):3773–3811, 2010.

[20] Gerard Hoberg and Gordon Phillips. Text-Based Network Industries and Endogenous Product Differentiation. *Journal of Political Economy*, 124(5):1423–1465, 2016.

[21] Gerard Hoberg and Gordon M. Phillips. Text-Based Industry Momentum. *Journal of Financial and Quantitative Analysis*, 53(6):2355–2388, 2018.

[22] Allen H. Huang, Hui Wang, and Yi Yang. FinBERT - A Large Language Model for Extracting Information from Financial Text, 2020.

[23] Jimin Huang, Mengxi Xiao, Dong Li, Zihao Jiang, Yuzhe Yang, Yifei Zhang, Lingfei Qian, Yan Wang, Xueqing Peng, Yang Ren, Ruoyu Xiang, Zhengyu Chen, Xiao Zhang, Yueru He, Weiguang Han, Shunian Chen, Lihang Shen, Daniel Kim, Yangyang Yu, Yupeng Cao, Zhiyang Deng, Haohang Li, Duanyu Feng, Yongfu Dai, VijayaSai Somasundaram, Peng Lu, Guojun Xiong, Zhiwei Liu, Zheheng Luo, Zhiyuan Yao, Ruey-Ling Weng, Meikang Qiu, Kaleb E. Smith, Honghai Yu, Yanzhao Lai, Min Peng, Jian-Yun Nie, Jordan W. Suchow, Xiao-Yang Liu, Benyou Wang, Alejandro Lopez-Lira, Qianqian Xie, Sophia Ananiadou, and Junichi Tsujii. Open-FinLLMs: Open Multimodal Large Language Models for Financial Applications, 2025.

[24] Junlin Julian Jiang and Xin Li. Look Ahead Text Understanding and LLM Stitching. *Proceedings of the International AAAI Conference on Web and Social Media*, 18:751–760, 2024.

[25] Alex Kim, Maximilian Muhn, and Valeri Nikolaev. Financial Statement Analysis with Large Language Models, 2024.

[26] Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R. Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, Yi Luan, Sai Meher Karthik Duddu, Gustavo Hernandez Abrego, Weiqiang Shi, Nithi Gupta, Aditya Kusupati, Prateek Jain, Siddhartha Reddy Jonnalagadda, Ming-Wei Chang, and Iftekhar Naim. Gecko: Versatile Text Embeddings Distilled from Large Language Models, 2024.

[27] Bradford Levy. Caution Ahead: Numerical Reasoning and Look-ahead Bias in AI Models, 2024.

[28] Jinchang Li, Ganghui Lian, and Aiting Xu. How do ESG affect the spillover of green innovation among peer firms? Mechanism discussion and performance study. *Journal of Business Research*, 158:113648, 2023.

[29] Lior Menzly and Oguzhan Ozbas. Market Segmentation and Cross-predictability of Returns. *The Journal of Finance*, 65(4):1555–1580, 2010.

[30] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space, 2013.

[31] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[32] Yuqi Nie, Yaxuan Kong, Xiaowen Dong, John M. Mulvey, H. Vincent Poor, Qingsong Wen, and Stefan Zohren. A Survey of Large Language Models for Financial Applications: Progress, Prospects and Challenges, 2024.

[33] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[34] Dhanesh Ramachandram and Graham W. Taylor. Deep Multimodal Learning: A Survey on Recent Advances and Trends. *IEEE Signal Processing Magazine*, 34(6):96–108, 2017.

[35] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, 2019.

[36] Nils Reimers and Iryna Gurevych. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation, 2020.

[37] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-Agnostic Interpretability of Machine Learning, 2016.

[38] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ”Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, pages 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[39] Suproteem K. Sarkar and Keyon Vafa. Lookahead Bias in Pretrained Language Models, 2024.

[40] Ramit Sawhney, Shivam Agarwal, Arnav Wadhwa, and Rajiv Ratn Shah. Deep Attentive Learning for Stock Movement Prediction From Social Media Text and Company Correlations. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8415–8426, Online, 2020. Association for Computational Linguistics.

[41] Jerick Shi and Burton Hollifield. Predictive Power of LLMs in Financial Markets, 2024.

[42] Yuxuan Tang and Zhanjun Liu. A Distributed Knowledge Distillation Framework for Financial Fraud Detection Based on Transformer. *IEEE Access*, 12:62899–62911, 2024.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[44] Haitao Wang, Jiale Zheng, Ivan E. Carvajal-Roca, Linghui Chen, and Mengqiu Bai. Financial Fraud Detection Based on Deep Learning: Towards Large-Scale Pre-training Transformer Models. In Haofen Wang, Xianpei Han, Ming Liu, Gong Cheng, Yongbin Liu, and Ningyu Zhang, editors, *Knowledge Graph and Semantic Computing: Knowledge Graph Empowers Artificial General Intelligence*, pages 163–177, Singapore, 2023. Springer Nature.

[45] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual E5 Text Embeddings: A Technical Report, 2024.

[46] Ziao Wang, Yuhang Li, Junda Wu, Jaehyeon Soon, and Xiaofeng Zhang. FinVis-GPT: A Multimodal Large Language Model for Financial Chart Analysis, 2023.

[47] Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780, 2022.

[48] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. BloombergGPT: A Large Language Model for Finance, 2023.

[49] Peng Xu, Xiatian Zhu, and David A. Clifton. Multimodal Learning With Transformers: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132, 2023.

[50] Yuhan Wang, Zhen Xu, Kunyuan Ma, Yuan Chen, and Jinsong Liu. Credit Default Prediction with Machine Learning: A Comparative Study and Interpretability Insights. 2024.

[51] Xuan Zhou and Yushen Huang. Unraveling Managerial Tangents in Firm Disclosure: Concealing Issues or Being Exposed?, 2023.