INTERCORPREL-LLM: ENHANCING FINANCIAL RELATIONAL UNDERSTANDING WITH GRAPH–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Identifying inter-firm relationships such as supply and competitive ties is critical for financial analysis and corporate governance, yet remains challenging due to the scale, sparsity, and contextual dependence of corporate data. Graph-based methods capture structure but miss semantic depth, while large language models (LLMs) excel at text but remain limited in their ability to represent relational dependencies. To address this, we propose InterCorpRel-LLM, a cross-modal framework that integrates GNNs with LLMs, supported by a proprietary dataset derived from FactSet supply chain records and three tailored training tasks: company graph matching, industry classification, and supply relation prediction. This design enables effective joint modeling of structure and semantics. Experiments show that InterCorpRel-LLM substantially outperforms strong baselines, including GPT-5, on a supply relation identification task, achieving an F-score of 0.8543 vs. 0.2287 with only a 7B-parameter backbone and lightweight training. The model also generalizes to zero-shot competitor identification, underscoring its ability to capture nuanced inter-firm dynamics. Our framework thus provides analysts and strategists with a robust tool for mapping and reasoning about complex corporate networks, enhancing decision-making and risk management in dynamic markets.

1 INTRODUCTION

Understanding inter-firm relationships—particularly supply dependencies and competitive dynamics—is central to modern finance and corporate strategy, as these ties shape resilience, performance, and systemic risk (Wei & Wang, 2010; Ivanov et al., 2025; Cao et al., 2025). Supply relations reveal the flow of goods and interdependencies across industries, where disruptions can cascade through networks and trigger financial shocks (Chen et al., 2022; Ivanov et al., 2025). Competitive linkages, meanwhile, influence strategic positioning and innovation diffusion; failing to monitor them risks "competitive blind spots" that leave firms exposed to unforeseen threats (Pant & Sheng, 2015). Accurately identifying such ties is therefore essential for financial analysis, risk management, and policy-making in an increasingly networked economy.

Modeling inter-firm relationships is challenging because the data are large-scale, sparse, and semantically rich. A single firm may connect to thousands of partners, yet networks are incomplete and context-dependent, with critical clues often buried in financial text or reports. Graph neural networks (GNNs) capture structural dependencies well but struggle when graphs are sparse or when meaning is encoded in language (Wasi et al., 2024; Kosasih & Brintrup, 2022). Conversely, large language models (LLMs) excel at interpreting nuanced text (Cao et al., 2025) but lack mechanisms to reason over graph structure. Thus, graph-only methods miss context, while text-only methods miss structure, leaving each insufficient on its own.

Recent work has explored GNN-LLM hybrids that integrate structure and semantics, but existing designs have clear limitations. LLM-centric approaches linearize graphs and lose topology, while GNN-centric approaches compress text and sacrifice nuance (Yang et al., 2024). Moreover, none have been adapted to the financial domain, where accurate modeling of supply chains and competition is essential. The absence of domain-specific frameworks and benchmarks has left analysts

without effective tools, underscoring the need for tailored graph–language models for inter-firm relationship analysis (Wasi et al., 2024).

To address this gap, we propose InterCorpRel-LLM, a GNN-LLM framework for modeling inter-firm relationships. Our method bridges structural and semantic modeling by combining graph data with domain-specific text. Using the FactSet Revere dataset, we construct a realistic domain-specific benchmark and design three complementary tasks: (1) Company Graph Matching to ground language in network context, (2) Industry Classification to capture high-level semantic groupings, and (3) Supply Relation Prediction to learn inter-company business relationships. Together, these tasks with lightweight training address core challenges in entity resolution, semantic understanding, and inter-firm relationship identification and enable the model to embed firms in a shared structural–textual space, capturing nuanced business relationships more effectively.

Our experiments show that InterCorpRel-LLM achieves substantial gains in both supply relation prediction and competitor identification, even in out-of-sample settings. In the supply relation task, it is significantly stronger than GPT-5, achieving an F-score of 0.8543 vs. 0.2287 with only a 7B-parameter backbone and lightweight training. In the competitor identification task, our dedicated training method greatly boosts the performance of the backbone model, enabling effective zero-shot detection of rival firms. Overall, InterCorpRel-LLM not only surpasses much larger open- and closed-source models, including GPT-5 and DeepSeek-v3.1, but also demonstrates that **carefully designed graph–language integration with domain-specific data and training**—rather than sheer parameter scale—is the key to robust inter-firm relationship modeling, with direct implications for financial analysis, strategy, and risk management.

2 RELATED WORKS

Supply Chain Relation Identification. Kosasih & Brintrup (2022) addressed supply chain opacity with a graph neural network (GNN)–based approach, relying solely on structural information while neglecting firm-specific context. Although effective on an automotive supplier dataset, their method has not been validated across broader industries. An alternative line of work leverages corporate reports or other public data to infer missing supply chain links. For example, Wichmann et al. (2020) applied natural language processing (NLP) techniques to online data, but their reliance on traditional models such as BiLSTM and SVM limited generalization and performance, particularly outside the automotive and aerospace sectors. More recently, Jin et al. (2025) introduced a paradigm that harnesses large language models (LLMs) to extract supply chain relations via contextual question–answering. This approach significantly improves generalization across industries, though it remains ineffective when firms deliberately conceal relationships absent from public sources.

Competitor Identification. Accurate recognition of rivals is central to valuation, governance, and strategic forecasting. Traditional approaches—such as industry codes (Phillips & Ormsby, 2016), managerial judgment, text-based similarity (Hoberg & Phillips, 2010; 2016), or curated databases—often suffer from incompleteness and "competitive blind spots" (Pant & Sheng, 2015). To overcome these limitations, Pant & Sheng (2015) introduced online isomorphism, showing that overlapping web content and hyperlink structures can effectively signal competition, outperforming offline baselines. More recently, Cao et al. (2025) demonstrated that large language models (LLMs) capture nuanced inter-firm similarities beyond conventional text-mining. Despite these advances, existing methods remain constrained in scalability, cross-industry generalization, and integration with structured financial networks. Developing approaches that jointly exploit network structure and semantic reasoning thus remains an open challenge for automated competitor identification.

GNN–LLM Integration. Recent studies have explored GNN–LLM integration in various domains. Wang et al. (2025) showed that pure LLMs underperform on graph tasks without explicit structure, motivating hybrid approaches. In recommendation field, Xi et al. (2024) proposed KAR to augment models with reasoning and factual knowledge from LLMs, while Wei et al. (2024) combined graph signals with LLM-based user–item representations. For open-ended graph reasoning, Tang et al. (2024) introduced GraphGPT with instruction tuning to align LLMs to graph structures, and Zhang et al. (2024) developed GraphTranslator to bridge pretrained GNNs with LLMs for both predefined and open-ended tasks, such as paper citation prediction and product recommendation. These advances highlight the promise of combining structural and semantic modeling, yet remain limited to several fields mentioned above. To our knowledge, we are the first to adapt a GNN–LLM frame-

work to inter-company relation modeling, addressing the unique challenges of supply chain and competitor analysis in finance.

All code and data-processing scripts will be released to ensure full reproducibility.

3 APPROACH

3.1 PRELIMINARIES

In this study, the task of identifying inter-firm business relationships is formulated based on the supply chain network of firms. Specifically, the inference of supply and competitive relations between firms is conducted with the known supply chain network as the foundation. The supply chain network of firms can be viewed as a special type of graph data. Compared to citation networks or product recommendation networks, supply chain data is enriched with more complex contextual information, potentially involving corporate operations, geopolitical conditions, and broader socioeconomic environments. This intrinSIC complexity fundamentally determines the difficulty of inferring inter-firm business relationships.

The identification of inter-firm supply relations can be regarded as a **directed edge prediction problem** in the supply chain network. The network is represented as

$$G = (V, E, X),$$

where V denotes the set of nodes, with each node representing a firm. The cardinality $|V| = N$ indicates the total number of firms. E denotes the set of directed edges in the network; if there exists an edge from node k to node g, this represents a supply relation where firm k provides certain products or services to firm g. The node feature matrix is denoted as $X \in \mathbb{R}^{N \times F}$, where each row corresponds to a firm and F is the dimension of the node feature vector. The supply relation prediction task is then defined as predicting the missing edges in E, given V, X, and the known subset of E.

To test whether our model truly understands the complex business relationships between enterprises, we also designed the task of identifying company competitors. The identification of inter-firm competitive relations can be formulated as a **binary classification task** grounded in information from the supply chain network. In this setting, all components of the supply chain graph $G = (V, E, X)$ are assumed to be known, and the goal is to predict whether a pair of firms forms a competitive relationship.

3.2 DATA ENCOMPASSING COMPREHENSIVE INFORMATION ON CORPORATE RELATIONSHIPS

To facilitate effective learning of inter-firm business relationships, we construct a domain-specific dataset from FactSet's Supply Chain Relationships records. We extract a directed graph of U.S. public companies from 2023, where nodes are firms and edges represent a known supplier→customer link (as documented by FactSet). The resulting supply network contains **3,211 firms spanning a wide range of sectors** and 11,635 verified inter-firm supply links. This graph provides the structural backbone of our data. FactSet's supply chain data is known to capture multi-tier supplier/customer networks and is used to uncover hidden dependency risks, making it an ideal foundation for our task.

We augment each company node with rich textual and categorical information to provide context that pure graph structure alone would miss. In particular, for each firm we include:

Business Description (Annual Report Text): Each firm's 2023 annual report (10-K filing) is processed to extract a synopsis of its core business activities, products, and financial highlights. These unstructured texts typically detail what the company does, its revenue streams, and its scale of operations. Such information is essential for understanding a firm's strategic positioning and the nature of its relationships.

Geographic Location: We attach the firm's headquarter location (country/region). Geographic context can be important, as supply chain ties often have regional patterns and risks (e.g. proximity can matter for certain logistics, and regulations/trade policies differ by region).

Industry Classification: Each firm is labeled with its industry category under the Standard Industrial Classification (SIC) system. SIC codes provide a hierarchical industry grouping (e.g. a firm

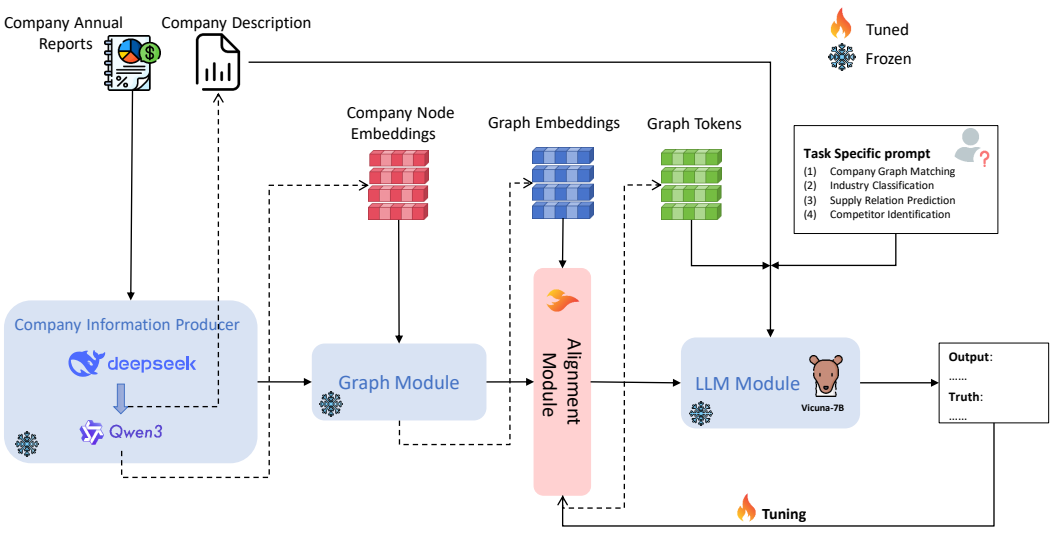

Figure 1: The overall structure of InterCorpRel-LLM

might be categorized as SIC 3674: Semiconductors and Related Devices). This gives the model a structured indication of the company's sector. Industry labels are valuable for relationship reasoning because, for instance, supply links typically connect firms in related industries (automotive manufacturers will link to auto parts suppliers, etc.), and competitors are usually in the same industry.

By combining the structural graph of who-supplies-whom with each firm's textual and categorical profile, our dataset provides a holistic view of the corporate ecosystem. This richness enables training tasks that require understanding not just network connections but also the business context behind those connections.

For competitor identification task, we obtain a competitor dataset from a commercial data provider that covers all firms in the above supply chain network and their pairwise competitive relations. This dataset enables us to evaluate whether the model can effectively identify competitors in a zero-shot setting, thereby testing its capacity to capture broader inter-firm dynamics.

3.3 InterCorpRel-LLM: Graph–Language Model for Company Relation Identification

Inspired by prior research(Tang et al., 2024; Zhang et al., 2024), the proposed framework consists of the following key components(Figure 1):

- (1) a *Company Information Producer*, which extracts and summarizes firm-level background information into vector representations;
- (2) a *Graph Module*, which captures structural information from the supply chain network based on the vector representations from *Company Information Producer*;
- (3) an *Alignment Module*, which maps graph-based embeddings into a representation space (graph tokens) more compatible with large language models (LLMs);
- (4) an *LLM Module*, which integrates graph-structured (graph tokens as part of the task prompt) and unstructured textual modalities for inter-company relation identification. For details on the prompt design, please refer to the attachment.

Company Information Producer. Leveraging LLMs to enhance data representations has been widely recognized as an effective approach for improving model performanceHong et al. (2025). In this study, we design a Company Information Producer that generates enriched node representations for each firm in the network. First, we employ the proprietary model DeepSeek-V3(Liu et al., 2024) to summarize annual reports of publicly listed firms (details provided in the Appendix).

Subsequently, we utilize the Qwen3-Embedding-8B model to produce 128-dimensional vector representations of firm nodes ($F = 128$). The Qwen3 series incorporates Multi-resolution Learning (MRL) during training, which ensures that truncating to the first n dimensions does not result in linear semantic loss; rather, the leading dimensions retain condensed semantic information(Zhang et al., 2025).

Graph Module. This module adopts a contrastive learning framework inspired by the CLIP (Contrastive Language–Image Pretraining) architecture, but applied to graph and text modalities. The objective is to align graph-based and text-based representations of the same firm while separating those of different firms. This encourages the firm embeddings to not only capture supply chain structural information but also reside closer to the textual embedding space. In this study, we employ the pretrained model released by Tang et al. (2024), who demonstrated that their model, trained on citation networks and product purchase data, effectively transmits graph information into embeddings that can be further processed by LLMs.

Alignment Module. To ensure that graph embeddings generated by the GNN module are effectively embedded into the LLM, we introduce an alignment layer between the Graph Model and the LLM Module, which maps the graph representations into a space more compatible with LLM inputs. Alignment Module is one fully connected layer.

LLM Module. Finally, the LLM integrates the graph-modality information from the Alignment Module with unstructured textual descriptions of firms to predict inter-firm relations. In this study, we employ `Vicuna-7B-v1.5` as the backbone LLM, consistent with numerous prior works on graph–language model integration(Cao et al., 2023; Wang et al., 2024; Wu et al., 2024).

3.4 Task Design for Corporate Relationship Understanding

We devise three focused training tasks, organized into a two-stage fine-tuning paradigm, to inject the graph-structured supply chain information into the LLM's representations. The goal is to enhance the model's semantic understanding of inter-firm relationships without altering the pre-trained GNN or LLM weights. Only a small Alignment Module (which maps graph embeddings to the LLM input space) is updated during training. This design keeps the majority of the model's knowledge intact while infusing new relational reasoning abilities.

Stage I: Company Graph Matching. The objective of this stage is to enable the LLM to better interpret supply chain structural signals processed by the Graph Module (GM) and Alignment Module. By training a lightweight cross-modal mapping, the model reduces its reliance on labeled data in later supervised tasks and achieves faster convergence. Specifically, we extract subgraphs centered on a target company with its 1-hop neighborhood and obtain vectorized representations using the GM. These are mapped via the Alignment Module into a sequence of graph tokens, with the number of tokens equal to the subgraph's node count. The task is formulated as a graph–text matching problem: aligning graph tokens with their corresponding firm textual identifiers (standardized company names). Formally, the probability of generating output O is:

$$P(O \mid S_H, S_T) = \prod_{i=1}^{L} p_\theta\big(x_i \mid S_H, S_{T,<i}, O_{<i}\big),$$

where S_H denotes the graph token sequence processed by the GM and Alignment modules, S_T is the company name sequence, L is the text length, and θ represents learnable parameters within the framework. Importantly, this stage is self-supervised (the "ground truth" name comes from the data itself), which helps the model bootstrap an alignment without needing labeled examples of supply links. After Stage I, the model has learned a preliminary joint space for graph and text: the graph-structured signals about a company and its neighbors start to influence the LLM's internal representations.

Stage II: Industry Classification and Supply Relation Prediction.

In the second stage, we jointly fine-tune the model on two supervised tasks that are highly relevant to business understanding: predicting a company's industry sector, and predicting supply chain links between companies. These tasks further refine the model's grasp of industry knowledge and relational logic.

Industry Classification Task: Each training instance consists of a single firm's subgraph (the firm plus its immediate neighbors in the supply chain network, as before) and the textual description of that firm (from its profile or annual report). The model is asked to output the firm's industry category (a specific SIC sector name). This is essentially a text classification task aided by graph context. The presence of certain neighbors in the graph can hint at the industry (for instance, if a company's neighbors include car parts suppliers and dealerships, the company is likely in the automotive sector). By learning this task, the model internalizes which features (both in text and graph) correlate with industry labels. This encourages the LLM to incorporate structural cues into its understanding of the firm's business domain.

Supply Relation Prediction Task: Here the model is given a pair of firms (A, B) along with their respective 1-hop neighbor subgraphs (excluding each other). The textual inputs include both firms' names and descriptions, plus the graph tokens from their subgraphs (via the Alignment Module). The task is to predict whether a directed supplier→customer relation exists from A to B. We formulate this as a binary classification or a conditional text generation (e.g. generating "Yes" if A supplies B, otherwise "No"). Successful prediction requires the model to analyze the compatibility of the two firms' profiles and their network contexts. For example, if A makes auto engines and B is an automobile manufacturer, and A is not already a neighbor of B, the model might predict a supply relation is plausible; conversely, if A is a retail store and B is a mining company with no common industry ground, it should predict no supply relation. Through this task, the model learns the structural logic of supply chains – which industry pairs typically form supplier-customer links, what patterns of shared neighbors or attributes indicate a likely connection, etc.

During Stage II, the model is optimized on both tasks. Notably, we keep the GNN and the core LLM frozen; only the Alignment Module is trainable. This strategy ensures we **preserve the pretrained knowledge** in the LLM (e.g. general language and commonsense, plus any financial knowledge it already has(You et al., 2025)) and in the GNN which could be pre-trained on generic graph structures, while just **infusing new domain-specific abilities**. It also avoids overfitting given our limited labeled data – by leveraging self-supervision (Stage I) and light supervised signals (Stage II), we achieve fast convergence with minimal data.

After fine-tuning, the unified model is expected to not only recall factual supply links, but also reason about them: it should infer likely relationships even between companies it has never seen before, based on their profiles and network context, and understand broader concepts like industry competition.

4 RESULT

4.1 EXPERIMENT SETUP

We evaluate our approach on two types of inter-firm relationship tasks: supply chain link prediction and competitor identification. To simulate a realistic scenario, we partition the supply chain graph data at the firm level with a 9:1 train–test split, ensuring that all companies in the test set are entirely unseen during training . For the supply link prediction task, we further divide test links into two scenarios(Zhang et al., 2022; Baek et al., 2020; Albooyeh et al., 2020):

(i) *Inductive / semi-inductive link prediction*, At least one endpoint of the link was seen during training. The inductive test set contains 1,954 positive supplier-customer links and 1,954 randomly sampled negative pairs (firm pairs with no supply relation).

(ii) *Fully inductive link prediction*, both firms in the link are unseen during training. This fully inductive test set is much smaller, with 95 positive links and 95 negatives.

This split allows us to assess generalization to completely new firms. In the fully inductive scenario, models cannot rely on any pre-existing supplier-customer link information for either node.

For the competitor relation prediction task, we use competition data recorded in a commercial dataset for the selected set of 3,211 firms in 2023, comprising a total of 8,895 competitive relations. From these, we randomly sample 2,000 positive relations and 2,000 negatives (non-existent competitor pairs) to construct the dataset. This task is used purely to test the zero-shot transferability of inter-firm reasoning learned from the supply chain domain.

4.2 BASELINE METHODS

We compare our proposed InterCorpRel-LLM framework against three categories of baseline methods:

(1) **Traditional GNN baselines:** We train graph neural networks on the observed supply chain graph structure. In particular, we use a Graph Attention Network (GAT)(Velickovic et al., 2017)and a GraphSAGE model(Hamilton et al., 2017). Both are enhanced with initial node embeddings derived from **Company Information Producer** that encodes firm background features into each node. This provides the GNNs some textual and attribute information to complement the graph topology.

(2) **Open-source LLMs:** We evaluate Vicuna-7B-v1.5-16k and Vicuna-13B-v1.5-16k, two open large language models, in a zero-shot manner. The prompt includes firm names, their business descriptions (from annual reports), geographic locations, and a structured context of their known suppliers/customers (formatted as a short text). We craft two prompt variants for each query: one without industry labels and one with industry labels (company's SIC sector name) appended to see if explicit industry context helps.

(3) **Closed-source LLMs:** We also test powerful proprietary models – DeepSeek-V3.1, GPT-4o, and GPT-5 – using the same zero-shot prompt templates (with and without SIC industry labels). These models represent state-of-the-art black-box LLMs and serve as an upper bound for language-only reasoning on our tasks.

In addition, we include ablations of our own model to gauge the impact of each training stage. Our full **InterCorpRel-LLM** is trained in two stages: Stage I aligns company names with the supply graph (via graph-text matching), and Stage II fine-tunes the model on the relational tasks. We consider two variants:

- **InterCorpRel-LLM_CGM**: A partial training baseline that is only trained on Stage I (Company Graph Matching) and not fine-tuned on any Stage II relational tasks.
- **InterCorpRel-LLM_IC_SRP**: Our model trained on Stage I, then Stage II with a multi-task objective combining Industry Classification(IC) and Supply Relation Prediction(SRP). This variant injects industry-related supervision during training in addition to the supply relation task.

4.3 PERFORMANCE IN SUPPLY CHAIN RELATIONSHIP IDENTIFICATION

Traditional graph neural networks (GraphSAGE and GAT) achieve moderately good accuracy on the inductive link prediction task (around 75–78% accuracy when at least one firm was seen in training). **However, their F1-scores are significantly lower, especially in the fully inductive scenario.**(Table 1) GAT's F1 drops to only 0.19 and SAGE's F1 drops to near 0 when predicting links between two completely unseen firms during training. This highlights the difficulty these structure-only models have in generalizing to entirely new nodes, even with company feature embeddings , a known weakness for GNNs in out-of-distribution settings. Overall, the GNN baselines can memorize and interpolate within the training graph, but they falter in the face of novel firms and links, as evidenced by their F1 on the fully inductive links task.

Zero-shot LLM baselines, despite their impressive general language understanding, underperform markedly on this structured link prediction task. Vicuna-7B and Vicuna-13B do show some capability on the inductive subset, for instance, Vicuna-7B achieves about 77% F1 inductively (without SIC labels). Yet, both Vicuna models degrade on the fully inductive set (Vicuna-7B F1 ≈ 0.59, and Vicuna-13B only 0.40). Interestingly, the 7B model outperforms the 13B model in our tests; the larger Vicuna may be overfitting to irrelevant textual patterns in the prompt instead of truly understanding the graph context, whereas the 7B model, being simpler, might generalize slightly better. The closed-source LLMs (DeepSeek-V3.1 and the GPT-4o/GPT-5 class models) fare even worse – they nearly fail completely to identify true supply links (e.g. DeepSeek-V3.1 achieves near-zero F1 on fully inductive links). Moreover, including the SIC industry labels in the prompt had mixed effects: in some cases it provided a marginal benefit, but often it made little difference or even confused the LLMs. For example, Vicuna-7B's inductive F1 decreased slightly from 0.7736 to 0.7111 when SIC codes were added. The inconsistent impact of adding industry label context suggests that

Table 1: Company supply relation prediction result

Model	without SIC label				with SIC label			
	Inductive		Fully inductive		Inductive		Fully inductive	
	Acc.	F-score	Acc.	F-score	Acc.	F-score	Acc.	F-score
GNN baselines								
GAT	0.7526	0.5359	0.6613	0.1887	–	–	–	–
SAGE	0.7774	0.5390	0.7253	0.0000	–	–	–	–
LLMs and our models								
vicuna-7b-v1.5	0.7377	0.7736	0.6579	0.5912	0.7326	0.7111	0.5895	0.4534
vicuna-13b-v1.6	0.6400	0.3973	0.6421	0.4040	0.6379	0.3891	0.6000	0.3117
deepseek-v3.1	0.5386	0.1375	0.5000	0.0000	0.5676	0.2216	0.5263	0.0957
GPT4o	0.5827	0.2680	0.5421	0.1458	0.0957	0.2692	0.5263	0.5827
GPT5	0.5548	0.1893	0.5263	0.0957	0.5660	0.2287	0.5316	0.1294
InterCorpRel-LLM_CGM	0.3930	0.5639	0.4053	0.5725	0.3984	0.5658	0.4105	0.5692
InterCorpRel-LLM_IC_SRP	**0.7968**	**0.8286**	**0.8105**	**0.8393**	**0.8347**	**0.8543**	**0.8368**	**0.8517**

Notes. "without/with SIC label" settings is only for LLMs and our models. "without SIC label" refers to predictions without industry code information, while "with SIC label" incorporates SIC labels. "Inductive" predicts links between a firm that was not observed during training and a firm that was included in the training data; "Fully inductive" predicts links between two firms that were both absent from the training phase. Following our evaluation protocol, F-score is the primary metric; accuracy is reported for completeness.

these off-the-shelf LLMs are not adept at incorporating structured business metadata when simply given as extra text in a prompt.

In stark contrast, our fine-tuned models achieve dramatically higher performance. The fully-trained InterCorpRel-LLM_IC_SRP attains an inductive F1 of 0.8543 and a fully-inductive F1 of 0.8517 with SIC label – the highest in each category by a large margin. And steady performance improvement can be found when the SIC information is added. **This indicates that our two-stage training effectively teaches the model how to reason about supplier relationships even for entirely new companies, while accurately leveraging industry SIC information when available and still maintaining strong performance in its absence.** Notably, the partially trained InterCorpRel-LLM_CGM model (Stage I only) performs significantly worse than the full model – its F1 scores hover around 0.56–0.57 in all cases, barely better than random guessing. This confirms that while Stage I is a crucial foundation, it is insufficient by itself for the complex task of predicting supply links. The model needs the Stage II task-specific fine-tuning to truly learn the semantics of inter-firm relationships.

These results clearly demonstrate that neither structure-only methods nor text-only methods are adequate for accurate supply chain link prediction in a sparse, dynamic business network. The **InterCorpRel-LLM** outperforms all baselines by a wide margin,especially in correctly identifying true supply relationships, highlighting the value of our hybrid approach. By integrating structural graph context into a language model and training with domain-specific objectives, the model learns to leverage both data modalities. The superior F1 scores of **InterCorpRel-LLM**, even on entirely new firms, underscore the necessity of domain-adaptive alignment – effectively teaching the model the "language" of company supply networks – to achieve high fidelity in inter-firm relationship modeling.

4.4 Performance in zero-shot competitor relationship identification

To test our model's generalization ability in identification of other business inter-firm relationships, we test **InterCorpRel-LLM** on a zero-shot task of identifying whether two firms are competitors, without any fine-tuning. In this task, the input given to the model is the same as that in the supply relation prediction task, including customized structure information and company text information

Table 2: Company competitor relation prediction result

Model	without SIC label		with SIC label	
	Acc.	F-score	Acc.	F-score
vicuna-7b-v1.5	0.7258	0.6347	0.7362	0.6535
vicuna-13b-v1.5	0.5895	0.5252	0.5805	0.4994
InterCorpRel-LLM_CGM	0.4150	0.5534	0.4150	0.5545
InterCorpRel-LLM_IC_SRP	**0.7855**	**0.7713**	**0.7965**	**0.7774**

Notes. "without SIC label" refers to predictions without industry code information, while "with SIC label" incorporates SIC labels. Following our evaluation protocol, F-score is the primary metric; accuracy is reported for completeness.

of the two focal companies. However, in this task, the model is required to judge whether there is a competitive relationship between the two focal copmanies. Considering Vicuna-7b is the backbone of our model and the better performance in the supply relation prediction task compared with closed-source LLMs, we only choose the Vicuna series model as our baseline in this task.

Table 2 shows Accuracy and F1-score on a balanced set of company pairs, comparing performance with and without SIC industry labels in the prompt. Despite no task-specific training, our **InterCorpRel-LLM** exhibit non-trivial skill in assessing competitive relationships.

Among open-source LLMs, Vicuna-7B gets a F1 of around 0.6535 and the 13B Vicuna gets a lower F1 around 0.5. Moreover, including the industry labels (SIC sector names) in the prompt gave the two model opposite marginal effects. Our **InterCorpRel-LLM_IC_SRP**, on the other hand, generalize remarkably well to this zero-shot competitor identification challenge, with an F1-score of 0.7774 on the balanced competitor dataset. Still, steady performance improvement can be found when the SIC information is added. For the Stage I-only model **InterCorpRel-LLM_CGM**, as expected, again underperforms on this task with a performance slightly above random.

This finding indicates that our model has internalized a robust conception of business competition, despite not being explicitly trained on a competitor identification task. It suggests that **InterCorpRel-LLM** has acquired generalizable reasoning patterns about inter-firm relationships through exposure to domain-specific data and carefully designed training objectives. Such a capability holds particular value for applications in market analysis and strategic intelligence.

5 Conclusion

Our experiments highlight the importance of integrating both textual and graph-based knowledge for modeling inter-firm relationships. The traditional GNN (structure-only) and LLM (text-only) baselines struggled to achieve high recall and precision on inter-firm relationship identification task, especially in out-of-distribution scenarios. In contrast, the **InterCorpRel-LLM** through its two-stage, multi-modal training – achieved significantly superior performance on supply relation prediction, and this strength generalized to an entirely new task of competitor identification. Our model's success underlines the value of domain-adaptive training signals: by aligning the language model with business-specific knowledge (like supply networks and industry categories), we enable it to make informed predictions about how companies relate to each other in the real world. The gains observed with InterCorpRel-LLM underscore the transferable and robust understanding that can be achieved by marrying corporate knowledge graphs with large language models, paving the way for more intelligent business analytics tools.

6 Code and Data Availability

The code, trained models, and data-processing scripts used in this paper will be released publicly upon acceptance. Due to licensing restrictions of FactSet data, we will publish sample data of annual report and company information. And there are some examples in the appendix.

References

Marjan Albooyeh, Rishab Goel, and Seyed Mehran Kazemi. Out-of-sample representation learning for multi-relational graphs. *arXiv preprint arXiv:2004.13230*, 2020.

Jinheon Baek, Dong Bok Lee, and Sung Ju Hwang. Learning to extrapolate knowledge: Transductive few-shot out-of-graph link prediction. *Advances in neural information processing systems*, 33:546–560, 2020.

He Cao, Zijing Liu, Xingyu Lu, Yuan Yao, and Yu Li. Instructmol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. *arXiv preprint arXiv:2311.16208*, 2023.

Yi Cao, Long Chen, Jennifer Wu Tucker, and Chi Wan. Can generative ai help identify peer firms? yi cao et al. *Review of Accounting Studies*, pp. 1–43, 2025.

Xia Chen, Guojin Gong, and Shuqing Luo. Short interest and corporate investment: evidence from supply chain partners. *Contemporary Accounting Research*, 39(2):1455–1508, 2022.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

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.

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

Mengze Hong, Wailing Ng, Chen Jason Zhang, Yifei Wang, Yuanfeng Song, and Di Jiang. Llm-in-the-loop: Replicating human insight with llms for better machine learning applications. *Authorea Preprints*, 2025.

Dmitry Ivanov, Alexandre Dolgui, Ajay Das, and Boris Sokolov. Digital supply chain twins: Managing the ripple effect, resilience, and disruption risks by data-driven optimization, simulation, and visibility. In *Handbook of Ripple Effects in the Supply Chain*, pp. 407–432. Springer, 2025.

Bohan Jin, Qianyou Sun, and Lihua Chen. Enhancing supply chain transparency in emerging economies using online contents and llms. In *2025 International Conference on Information Networking (ICOIN)*, pp. 487–492. IEEE, 2025.

Edward Elson Kosasih and Alexandra Brintrup. A machine learning approach for predicting hidden links in supply chain with graph neural networks. *International Journal of Production Research*, 60(17):5380–5393, 2022.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Gautam Pant and Olivia RL Sheng. Web footprints of firms: Using online isomorphism for competitor identification. *Information Systems Research*, 26(1):188–209, 2015.

Ryan L Phillips and Rita Ormsby. Industry classification schemes: An analysis and review. *Journal of Business & Finance Librarianship*, 21(1):1–25, 2016.

Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin Su, Suqi Cheng, Dawei Yin, and Chao Huang. Graphgpt: Graph instruction tuning for large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 491–500, 2024.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. Graph attention networks. *stat*, 1050(20):10–48550, 2017.

Duo Wang, Yuan Zuo, Fengzhi Li, and Junjie Wu. Llms as zero-shot graph learners: Alignment of gnn representations with llm token embeddings. *Advances in Neural Information Processing Systems*, 37:5950–5973, 2024.

Yuxiang Wang, Xinnan Dai, Wenqi Fan, and Yao Ma. Exploring graph tasks with pure llms: A comprehensive benchmark and investigation. *arXiv preprint arXiv:2502.18771*, 2025.

Azmine Toushik Wasi, MD Islam, Adipto Raihan Akib, and Mahathir Mohammad Bappy. Graph neural networks in supply chain analytics and optimization: Concepts, perspectives, dataset and benchmarks. *arXiv preprint arXiv:2411.08550*, 2024.

Hsiao-Lan Wei and Eric TG Wang. The strategic value of supply chain visibility: increasing the ability to reconfigure. *European Journal of Information Systems*, 19(2):238–249, 2010.

Wei Wei, Xubin Ren, Jiabin Tang, Qinyong Wang, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. Llmrec: Large language models with graph augmentation for recommendation. In *Proceedings of the 17th ACM international conference on web search and data mining*, pp. 806–815, 2024.

Pascal Wichmann, Alexandra Brintrup, Simon Baker, Philip Woodall, and Duncan McFarlane. Extracting supply chain maps from news articles using deep neural networks. *International Journal of Production Research*, 58(17):5320–5336, 2020.

Yuxia Wu, Shujie Li, Yuan Fang, and Chuan Shi. Exploring the potential of large language models for heterophilic graphs. *arXiv preprint arXiv:2408.14134*, 2024.

Yunjia Xi, Weiwen Liu, Jianghao Lin, Xiaoling Cai, Hong Zhu, Jieming Zhu, Bo Chen, Ruiming Tang, Weinan Zhang, and Yong Yu. Towards open-world recommendation with knowledge augmentation from large language models. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pp. 12–22, 2024.

Haotong Yang, Xiyuan Wang, Qian Tao, Shuxian Hu, Zhouchen Lin, and Muhan Zhang. Gl-fusion: Rethinking the combination of graph neural network and large language model. *arXiv preprint arXiv:2412.06849*, 2024.

Yuxin You, Zhen Liu, Xiangchao Wen, Yongtao Zhang, and Wei Ai. Large language models meet graph neural networks: a perspective of graph mining. *Mathematics*, 13(7):1147, 2025.

Daokun Zhang, Jie Yin, and S Yu Philip. Link prediction with contextualized self-supervision. *IEEE transactions on knowledge and data engineering*, 35(7):7138–7151, 2022.

Mengmei Zhang, Mingwei Sun, Peng Wang, Shen Fan, Yanhu Mo, Xiaoxiao Xu, Hong Liu, Cheng Yang, and Chuan Shi. Graphtranslator: Aligning graph model to large language model for open-ended tasks. In *Proceedings of the ACM Web Conference 2024*, pp. 1003–1014, 2024.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

A Appendix

A.1 Company Data Sample

Table 3: Company Data Sample

Year	Name	Country	Province	CIK	SIC_Label
2023	Tejon Ranch Co.	US	California	96869	Crops
2023	A. O. Smith Corp.	US	Wisconsin	91142	Household Appliances
2023	Teleflex, Inc.	US	Pennsylvania	96943	Surgical and Medical Instruments and Apparatus
2023	Telephone & Data Systems, Inc.	US	Illinois	1051512	Radiotelephone Communications
2023	TTEC Holdings, Inc.	US	Texas	1013880	Business Services, Not Elsewhere Classified
2023	TESSCO Technologies, Inc.	US	Maryland	927355	Electronic Parts and Equipment, Not Elsewhere Classified
2023	Tetra Tech, Inc.	US	California	831641	Engineering Services
2023	TETRA Technologies, Inc.	US	Texas	844965	Oil and Gas Field Services, Not Elsewhere Classified
2023	Texas Instruments Incorporated	US	Texas	97476	Semiconductors and Related Devices
2023	Textron, Inc.	US	Rhode Island	217346	Aircraft
2023	The Timken Co.	US	Ohio	98362	Ball and Roller Bearings
2023	VOXX International Corp.	US	Florida	807707	Household Audio and Video Equipment
2023	Thermo Fisher Scientific, Inc.	US	Massachusetts	977475	Laboratory Analytical Instruments
2023	Aziyo Biologics, Inc.	US	Maryland	1708527	Biological Products, except Diagnostic Substances
2023	Kadant Inc.	US	Massachusetts	886346	Special Industry Machinery, except Metalworking
2023	The Charles Schwab Corp.	US	Texas	316709	Investment Advice
2023	Thor Industries, Inc.	US	Indiana	730263	Miscellaneous Transportation Equipment
2023	Tidewater, Inc.	US	Texas	98222	Water Transportation

A.2 COMPANY DESCRIPTION SAMPLE

The company descriptions following are all result of original company annual reports processed by Deepseek-V3.

MongoDB, Inc.: MongoDB, Inc. is a leading developer data platform company that empowers organizations to innovate through software and data. The company offers an integrated suite of database and related services designed to support modern application development across various industries and use cases. MongoDB's flagship product is its document-based database, which combines the flexibility of non-relational databases with key features of traditional relational databases, enabling developers to build and scale applications efficiently.

The company operates primarily through two key offerings: **MongoDB Atlas**, a fully managed multi-cloud database-as-a-service (DBaaS) solution, and **MongoDB Enterprise Advanced**, a self-managed commercial database for enterprise deployments. MongoDB Atlas provides automated provisioning, monitoring, and security, allowing customers to focus on application development rather than infrastructure management. Enterprise Advanced caters to organizations requiring on-premises, hybrid, or customized cloud deployments with advanced security and management features.

MongoDB serves a diverse customer base, including enterprises across industries such as financial services, healthcare, retail, and technology. The company follows a developer-centric approach, fostering a large global community through free offerings like **Community Server** and a **MongoDB Atlas free tier**, which help drive adoption and eventual conversion to paid subscriptions.

With a strong focus on innovation, MongoDB continues to expand its platform with additional capabilities such as **search, time-series data handling, mobile synchronization, and analytics integrations**, reducing the need for multiple specialized database solutions. The company operates globally, with significant revenue generated outside the U.S., and maintains strategic partnerships with major cloud providers (AWS, Google Cloud, and Microsoft Azure) and system integrators to enhance market reach and customer adoption.

MongoDB's growth strategy includes acquiring new customers, expanding within existing accounts, extending product leadership, and strengthening its developer community and partner ecosystem. The company is publicly traded on Nasdaq under the symbol **MDB**.

Amplify Energy Corp.: Amplify Energy Corp. is an independent oil and natural gas company engaged in the acquisition, development, and production of oil and natural gas properties across key U.S. regions, including Oklahoma, the Rockies (Bairoil), federal waters offshore Southern California (Beta), East Texas/North Louisiana, and the Eagle Ford. The company operates primarily in mature oil and gas reservoirs, focusing on both operated and non-operated working interests in producing and undeveloped leasehold acreage.

As of December 31, 2022, Amplify Energy reported total estimated proved reserves of approximately 124.0 million barrels of oil equivalent (MMBoe), with a significant portion classified as proved developed reserves. The company's production mix includes natural gas (42%), oil (39%), and natural gas liquids (NGLs) (19%). Amplify operates a substantial portion of its assets, managing properties that account for 92% of its total proved reserves.

Key operational segments include: - **Oklahoma**: Focused on wells in Alfalfa and Woods counties, contributing 28% of proved reserves. - **Rockies (Bairoil)**: Primarily located in Wyoming's Lost Soldier and Wertz fields, accounting for 23% of reserves. - **Southern California (Beta)**: Offshore production platforms (Ellen, Eureka, and Elly) with a 16-inch pipeline, contributing 11% of reserves (currently non-producing due to a pipeline incident in October 2021). - **East Texas/North Louisiana**: Includes fields such as Joaquin and Carthage, representing 35% of reserves. - **Eagle Ford (Non-Op)**: Non-operated assets in the Eagleville fields, making up 2% of reserves.

Amplify Energy markets its production under month-to-month contracts with major customers, including HF Sinclair Corporation, Southwest Energy LP, and Koch Energy Services. The company relies on third-party midstream services for NGL commitments in Oklahoma.

The company's operations are subject to commodity price volatility, regulatory oversight, and operational risks, including the recent pipeline incident in Southern California. Amplify Energy continues

to focus on reserve replacement, cost management, and strategic development to sustain production and financial performance.

A.3 EXPERIMENT PROMPTS

Prompt for extracting key information from corporate annual reports :

You are a professional financial analysis assistant. Based on the provided information on the company, generate an English business description that describes the main business model, the segments the company operates in and the products the company offers. The description should be written from an outsider's perspective. Do not use other information you may have on the company. The description should not exceed 200 tokens. Just provide the description, do not add further comments. Please summarize the following important information in the company's annual report:{text}.

Prompt for Company Graph Matching Task for InterCorpRel-LLM *(<graph>is a placeholder that will be replaced by graph tokens generated by the Alignment Module. {company_names} is the text sequence of company names to be matched.)* :

Given a sequence of graph tokens <graph>that constitute a subgraph of an industry supply chain graph, where the first token represents the central node of the subgraph, and the remaining nodes represent the first order neighbors of the central node. Each graph token contains the content of the introduction of the company. If a company supply some product or some service to another one, a link is constructed from the former company to the other one. Here is a name list of companies:{company_names}, please reorder the name list according to the order of graph tokens (i.e., complete the matching of graph tokens and company-name list).

Prompt for Industry Classification Task for InterCorpRel-LLM *(<graph>is a placeholder that will be replaced by graph tokens generated by the Alignment Module. {company_description} is the focal company textual description. {company_name} is the the focal company's name.)*:

Given an industry supply chain graph: <graph>where the 0th node is the central company, and other nodes are its one-hop or multi-hop neighbors, with the following information: Description: { company_description } Company name: { company_name } Question: Which industry category does the company belong to under the SIC classification system? Give the most likely SIC industry category of this company directly, in the form "full name of the category" .

Prompt for Supply Relation Prediction Task for InterCorpRel-LLM *(<graph1/2>is a placeholder that will be replaced by graph tokens generated by the Alignment Module. {company1/2_description} is the focal company textual description. {company1/2_name} is the the focal company's name.)*:

Given a sequence of graph tokens: <graph1>that constitute a subgraph of an industry supply chain graph, where the first token represents the central node of the subgraph, and the remaining nodes represent the first order neighbors of the central node. The information of the central node is as follow: Description: {company1_description} Company name: {company1_name}. and the other sequence of graph tokens: <graph2>, where the first token (the central node) with the following information: Description: {company2_description}. Company name: {company2_name}. If the link from node 1 to node 2 represent the supply chain relationship from the former company to the latter company, is there a link from node 1 to node 2? Give me a direct answer of "yes" or "no".

Prompt for Competitor Identification Task for InterCorpRel-LLM *(<graph1/2>is a placeholder that will be replaced by graph tokens generated by the Alignment Module. {company1/2_description} is the focal company textual description. {company1/2_name} is the the focal company's name.)*:

Given a sequence of graph tokens: <graph1>that constitute a subgraph of an industry supply chain graph, where the first token represents the central node of the subgraph, and the remaining nodes represent the first order neighbors of the central node. The information of the central node is as follow: Description: {company1_description} Company name: {company1_name}. and the other sequence of graph tokens: <graph2>, where the first token (the central node) with the following information: Description: {company2_description}. Company name: {company2_name} . If the link of the industry supply chain graph represents a supply chain relationship from the source com-

pany to the target company, whether the two companies represented by the central nodes of the two subgraphs are competitors of each other in business? Give me a direct answer of "yes" or "no".

Prompt for Supply Relation Prediction Task for LLMs *({company1/2_description} is the focal company textual description. {company1/2_name} is the the focal company's name. {company1/2_connections} is a text description of the focal company's first-tier suppliers and first-tier customers.)*:

Given two graph nodes which are both subgraphs from a graph of an industry supply chain graph. The information of the first central node is as follow: Description: {company1_description} Company name: {company1_name}. Connections: {company1_connections} and the second central node with the following information: Description: {company2_ description}. Company name: {company2_name}. Connections: {company2_connections} If the link from node 1 to node 2 represent the supply chain relationship from the former company to the latter company, Is there a link from node 1 to node 2? Give me a direct answer of "yes" or "no".

Prompt for Competitor Identification Task for LLMs *({company1/2_description} is the focal company textual description. {company1/2_name} is the the focal company's name. {company1/2_connections} is a text description of the focal company's first-tier suppliers and first-tier customers.)*:

Given two sequences of graph nodes which are both subgraphs from a graph of an industry supply chain. The information of the first central node in the first subgraph is as follow: Description: {company1_description} Company name: {company1_name}. Connections: {company1_connections} and the second central node in the other subgraph with the following information: Description: {company2_description} Company name: {company2_name}. Connections: {company2_connections} If the link of the industry supply chain graph represents a supply chain relationship from the source company to the target company, whether the two companies represented by the core nodes of the two subgraphs are competitors of each other in business? Give me a direct answer of "yes" or "no".

A.4 LLM USAGE

LLM is only used to aid or polish writing in this paper. Specifically, LLM is used to check spelling and grammar, and to polish the expression of certain sentences.