Fusing LLMs and KGs for Formal Causal Reasoning behind Financial Risk Contagion

Guanyuan Yu, Xv Wang, *Qing Li *Member, IEEE*, Yu Zhao *Member, IEEE*

Abstract—Financial risks trend to spread from one entity to another, ultimately leading to systemic risks. The key to preventing such risks lies in understanding the causal chains behind risk contagion. Despite this, prevailing approaches primarily emphasize identifying risks, overlooking the underlying causal analysis of risk. To address such an issue, we propose a **R**isk **C**ontagion **C**ausal **R**easoning model called **RC²R**, which uses the logical reasoning capabilities of large language models (LLMs) to dissect the causal mechanisms of risk contagion grounded in the factual and expert knowledge embedded within financial knowledge graphs (KGs). At the data level, we utilize financial KGs to construct causal instructions, empowering LLMs to perform formal causal reasoning on risk propagation and tackle the "causal parrot" problem of LLMs. In terms of model architecture, we integrate a fusion module that aligns tokens and nodes across various granularities via multi-scale contrastive learning, followed by the amalgamation of textual and graph-structured data through soft prompt with cross multi-head attention mechanisms. To quantify risk contagion, we introduce a risk pathway inference module for calculating risk scores for each node in the graph. Finally, we visualize the risk contagion pathways and their intensities using Sankey diagrams, providing detailed causal explanations. Comprehensive experiments on financial KGs and supply chain datasets demonstrate that our model outperforms several state-of-the-art models in prediction performance and out-of-distribution (OOD) generalization capabilities. We will make our dataset and code publicly accessible to encourage further research and development in this field.

Index Terms—Large Language Models, Financial Knowledge Graphs, Causal Reasoning, Financial Risk Contagion.

1 Introduction

Financial risk contagion, the phenomenon where the risk from one financial entity rapidly spreads to others, can escalate into systemic risks if not properly controlled [1], [2], [3], [4]. The essence of financial risk control lies in a profound understanding of the causal mechanisms behind risk contagion. As shown in Fig. 1(a), through formal causal inference, we can ensure that the causal chain of risk contagion is $A \rightarrow B \rightarrow C$. Consequently, we can design effective strategies to block the chain of risk contagion. However, the majority of existing studies concentrate on the development of risk recognition models based on machine learning [5], [6], [7], [8], [9], [10] and deep learning [11], [12], [13], [14], [15], [16], [17], [18]. They neglect the intricate causal mechanisms behind financial risk contagion, hindering the implementation of risk prevention and control strategies.

To address such issues, this study proposes a novel **R**isk **C**ontagion **C**ausal **R**easoning model called **RC²R**, which uses the logical reasoning capabilities of large language models (LLMs) [19], [20], [21], [22] to analyze the causal mechanisms behind risk contagion, based on factual and professional knowledge in the financial knowledge graphs (KGs) [23], [24]. Specifically, as shown in Fig. 1(a), we first perform random interventions on the financial KGs, and then guide the LLMs to conduct formal causal reasoning to identify the risk contagion paths. Based on information retrieved from KGs, LLMs generate explanations for queries
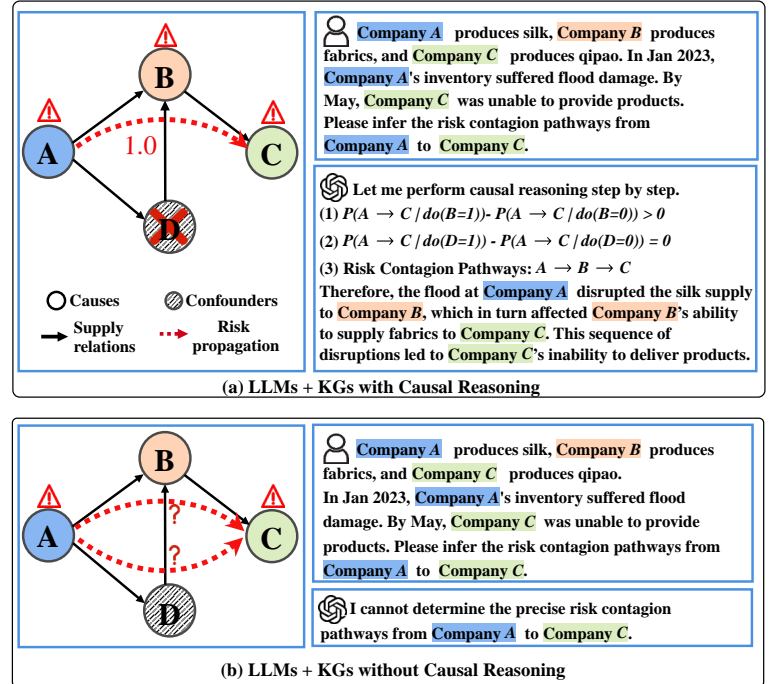

Fig. 1: A Case of Risk Causal Reasoning. (a) It utilizes formal causal inference to eliminate confounders as well as accurately trace and quantify the pathways of risk contagion, whereas (b) fails to accomplish this.

while quantifying the intensity of risk propagation. For instance, to determine if node B is part of the financial risk contagion path, we verify the change in contagion probability before and after intervening at node B. A greater change suggests a higher likelihood of node B being on the contagion path, while a smaller change indicates its lesser involvement.

Notably, the data foundation for LLMs' inference is natural language, whereas financial KGs are structured as graphs, leading to significant disparities in data modality. Here, we seamlessly integrate these two modalities from two perspectives. **(i) At the data level**, to guide LLMs in performing formal causal reasoning, we develop causal instructions. Such instructions include prompts, queries, and explanations, all based on financial KGs. Both KGs and instructions are then inputted into GNNs and LLMs, generating high-level representations. **(ii) At the architectural level**, we introduce a fusion module that employs a multi-scale contrastive loss function to explicitly coordinate the alignments between textual data and financial KGs. This module then leverages soft prompt with cross multi-head attention mechanisms to facilitate interaction from LLMs to KGs. Following this, a risk pathway inference module is implemented to identify and analyze predominant risk propagation pathways effectively.

During the inference stage, LLMs trace the pathways of financial risk contagion on KGs according to queries, generating explanations. Then, we can utilize Sankey diagrams to visualize the direction and intensity of risk contagion. In essence, we employ LLMs as a causal reasoning engine to explore risk contagion paths within financial KGs. In summary, our study makes the following three unique contributions.

• To the best of our knowledge, this study represents the first effort to activate the formal causal reasoning capabilities of LLMs while utilizing the factual and specialized knowledge embedded in financial KGs. This approach provides profound insights into the causal mechanisms of risk contagion and opens up innovative methodologies and perspectives for financial risk analysis.

• We introduce multi-scale contrastive learning to align tokens and nodes, and employ soft prompt with cross multi-head attention to seamlessly integrate text and graph information. In the risk pathway inference module, we introduce intervention mechanisms to precisely estimate the impact of each variable on risk contagion.

• Our model demonstrates excellent performance and successfully visualizes the pathways of risk contagion. We have made our datasets and code publicly available for further research and development[1]. To our knowledge, we are the first to open-source comprehensive datasets for causal analysis of financial risk contagion, including causal corpora and financial KGs.

2 RELATED WORK

2.1 Financial Risk Detection & Analysis

Current studies usually leverage deep learning models for risk detection and analysis. For example, many unified

[1]. Please refer to https://github.com/wang1595243339/RC2R.

models based on the convolutional neural networks (CNNs) and long short-term memory (LSTM) are introduced to process financial data, providing insights into market fluctuations [25], [26], [27], [28]. The graph neural network (GNN) family (e.g., Risk-Rate [29], Know-GNN [30], MAGNN [18], DGA-GNN [31]) is utilized to analyze trading data within financial networks, identifying unusual transactions and financial risks. Besides, Transformer and its variants (e.g., Html [32], Numhtml [33], FinBERT [34], GPT-3 [35]) are applied to analyze financial risks. However, these models often overlook the complex causal mechanisms behind financial risk contagion, hindering effective risk prevention and control strategies. To address this issue, we propose integrating LLMs and KGs to investigate the inherent causal mechanisms of risk contagion.

2.2 Causal Inference

Classical studies have utilized deep learning models to identify causal structures and separate causal variables [36], [37]. For instance, deep structural causal models (DSCMs) employ normalizing flows and variational inference for tractable inference of exogenous noise variables [38]. Techniques like DIR [39] and V-REx [40] are used to differentiate causal from non-causal features in input data.

Nowadays, more and more studies concentrate on using LLMs to discover causal relationships and estimate causal effects [41], [42], [43], [44]. For example, GPT family (e.g., GPT-3 [45], ChatGPT [46]) is used to answer causal discovery questions and look for causal directions. DISCO makes the estimation of causal effects by generating counterfactual data [47]. CInA performs self-supervised causal learning and facilitates zero-shot causal inference [48].

As an advancement of the above studies, we are the first to explore causal discovery and estimation in financial risk contagion by integrating LLMs with KGs. Our approach provides deep insights into financial risk contagion and aids in the development of risk prevention and control strategies.

2.3 Fusion of LLMs and KGs

LLMs have been shown to exhibit hallucinations [49]. To address this issue, some studies have focused on integrating the specialized and factual knowledge of KGs into the inference process of LLMs [24], [50]. Broadly, there are two popular approaches to this integration.

The first approach treats LLMs as intermediary agents that interact with KGs [51]. For instance, StructGPT [52] introduces special interfaces that enable LLMs to access and reason with information from KGs. Similarly, Think-on-graph [53] identifies reasoning paths to elucidate how LLMs infer and generate answers.

The second approach augments LLMs with KG knowledge during the training or fine-tuning phases. This typically involves using separate encoders for text and graph-based data, as seen in models like KagNet [54] and QA-GNN [55]. JointLK [56] further refines this method by scoring reasoning results at each step through a bi-directional attention mechanism.

In our study, to mitigate hallucinations in LLMs during causal reasoning, we enhance the integration of LLMs and KGs by employing a multi-scale contrastive loss function
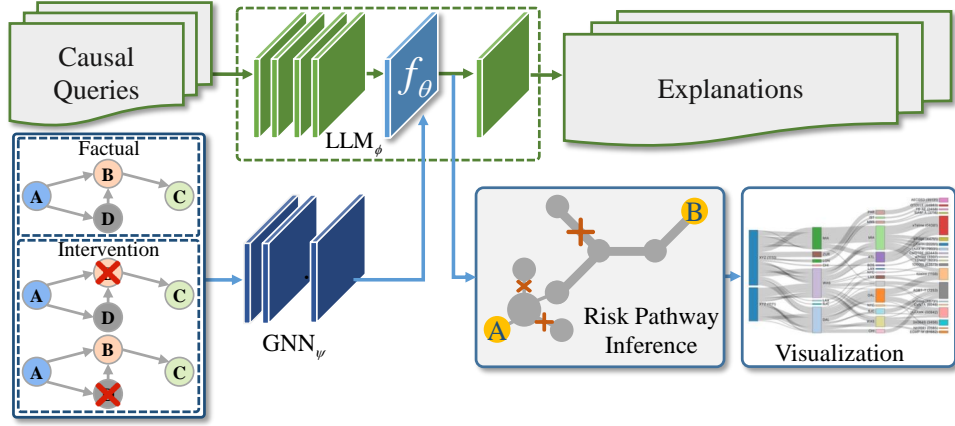
Fig. 2: The Whole Framework of Our Proposed RC²R

and soft prompts with cross multi-head attention mechanisms.

3 OUR APPROACH

Fig. 2 provides an overview of RC²R. Initially, casual queries and financial KGs are processed through a large language model $\text{LLM}_\phi(\cdot)$ and a graph neural network $\text{GNN}_\psi(\cdot)$, respectively, to produce high-level representations. These representations are then integrated in a latent space by a fusion module, denoted as $f_\theta(\cdot)$. Following this integration, $\text{LLM}_\phi(\cdot)$ leverages such representations to generate responses and corresponding explanations. Additionally, a risk pathway inference module is employed to identify the predominant propagation pathways. To depict the direction and magnitude of risk contagion, Sankey diagrams are finally utilized.

Notably, the key to achieving the aforementioned objective lies in the application of LLMs to accurately identify the ground-truth causal relationships of risk propagation within financial KGs. Nevertheless, LLMs can generate text that appears coherent and logically sound, effectively mimicking causal reasoning without truly grasping the underlying causal relationships. Such responses are based on pattern recognition from extensive training datasets rather than an actual understanding of real-world dynamics or independent logical reasoning, that is, LLMs may be "causal parrots" [57], [58].

For the first time, we innovatively employ the following three steps to pioneer the activation of causal reasoning capabilities in LLMs. (i) We develop a formal causal diagram (Fig. 3) to understand, analyze, and predict the causal relationships of risk contagion. (ii) In accordance with well-defined formal rules of causal inference, we formulate various causal queries and provide detailed explanations to guide LLMs in recognizing cause-and-effect relationships in risk contagion, as depicted in Table 1. (iii) We seamlessly integrate LLMs with GNNs via our proposed fusion module.

3.1 Formal Causal Reasoning for Risk Contagion

Leveraging the inference capabilities of LLMs to conduct an in-depth analysis of risk propagation mechanisms in financial KGs primarily involves identifying causal variables within these graphs. Here, let $\mathcal{G} = \{X, Z\}$ denote the KGs, where X represents the causal components, i.e., factors influencing risk contagion, and Z signifies the non-causal components, i.e., factors not affecting risk contagion. Meanwhile, Y is the outcome variable that reflects the result of risk propagation. Here, we utilize the Structural Causal Model (SCM) [59] to thoroughly explore the causal relations between these variables. Fig. 3 illustrates two types of relations: (i) $X \to Y$ means that the casual part X is the endogenous parent responsible for determining Y. Taking the risk contagion in Fig. 1(a) as an example, *Company A → Company B → Company C* is the casual part, which perfectly explains the mechanism of risk contagion. (ii) $X \leftarrow Z \to Y$ reflects that Z is a confounder between X and Y, which opens a backdoor path. To discern the ground-truth causal relationships of risk contagion, we need to block such a backdoor pathway.

Fig. 3: Visualization of SCM via a Directed Acyclic Graph

Here, we attempt to integrate LLMs and financial KGs to block the backdoor paths. Specifically,

$$Y = f_Y(X), Y \perp\!\!\!\perp Z \mid X,$$
$$s.t. \ X = f_\theta(\text{LLM}_\phi(\mathcal{I}), \text{GNN}_\psi(\mathcal{G}_{\text{set}})).$$

In the above equation, $\text{LLM}_\phi(\cdot)$ represents a large language model equipped with a parameter set denoted by ϕ. $\text{GNN}_\psi(\cdot)$ refers to a graph neural network characterized by its parameter set ψ. $\mathcal{I}$ corresponds to the causal instruction, depicted in Table 1. $\mathcal{G}_{\text{set}}$ denotes the set of factual and intervention graphs in Algorithm 1. The function $f_\theta(\cdot, \cdot)$ is responsible for integrating the instruction information and graph information within a latent space.

TABLE 1: Causal Instruction & Data Invervention

Role: You are a professional assistant specializing in the causal reasoning of financial risk contagion. Please use your knowledge of financial risk to answer the following questions.
Query: *Company A* produces silk, *Company B* produces fabrics, and *Company C* produces qipao. In Jan 2023, *Company A*'s inventory suffered flood damage. By May, *Company C* was unable to provide products. Please infer the risk contagion pathways from *Company A* to *Company C*.

Groups	Graphs	Contagion Probabilities
Factual[*]		$P(A \rightarrow C\|do(A = 1)) = 1$ $P(A \rightarrow C\|do(B = 1)) = 1$ $P(A \rightarrow C\|do(D = 1)) = 1$
Intervention		$P(A \rightarrow C\|do(A = 0)) = 0$
		$P(A \rightarrow C\|do(B = 0)) = 0$
		$P(A \rightarrow C\|do(D = 0)) = 1$

Explanation: Let me perform formal causal reasoning step by step: (1) $P(A \rightarrow C\|do(A = 1)) - P(A \rightarrow C\|do(A = 0)) > 0$, (2) $P(A \rightarrow C\|do(B = 1)) - P(A \rightarrow C\|do(B = 0)) > 0$, (3) $P(A \rightarrow C\|do(D = 1)) - P(A \rightarrow C\|do(D = 0)) = 0$, (4) The causal chains of risk propagation is $A \rightarrow B \rightarrow C$. The flood at *Company A* disrupted the silk supply to *Company B*, which in turn affected *Company B*'s ability to supply fabrics to *Company C*. This sequence of disruptions led to *Company C*'s delivery delays.

[*] Notably, the nodes in financial KGs precisely correspond to the entities in the text. Here, we visually demonstrate their matching relations through colors. In the subsequent fusion module, we continue to maintain the correct alignments between such two features.

3.2 Casual Instruction & Data Intervention

Definition 1 (Hierarchies of Causation). Proposed by Pearl and Mackenzie [60], it is a three-tiered framework for causal reasoning, including association, intervention, and counterfactual. This framework is pivotal for a deep understanding and accurate inference of causal relationships.

● **Association:** This focuses on statistical dependencies among random variables, employing probabilistic reasoning regarding joint and conditional distributions, that is $P(X = x, Y = y)$ and $P(Y = y|X = x)$.

● **Intervention:** This allows us to artificially adjust the value of variables (causes) within a causal system, thereby observing the impact of this adjustment on other variables (effects). Because it can control for potential confounding variables, intervention enables us to more directly test and validate causal hypotheses.

Such interventions can be formalized using the do-operator [61], expressed as the distribution of Y being $P(Y = y|do(X = x))$ when setting $X = x$.

● **Counterfactual:** This is a hypothetical way of thinking that explores the impact on outcomes by imagining events as not having occurred or as different from reality, aiming to answer the question, "What would the outcome be if things were different?" Counterfactual probabilities can be written as $P(Y_x = y)$, representing the probability that "Y would be y, had X been x".

Based on the above hierarchies of causation, we construct two categories of data (as depicted in Table 1) to block backdoor pathways and fine-tune our model to understand the cause-and-effect relationships behind risk contagion.

Algorithm 1: Data Intervention Algorithm

Input: Financial knowledge graph (KG), start node v_s, target node v_t.
Output: Set of factual and intervention graphs $\mathcal{G}_{\text{set}}$.

1   $\mathcal{G}_{\text{set}} \leftarrow \emptyset$;
    // Perform a depth-first search (DFS) to obtain the factual graph $\mathcal{G}$ from the financial KG
2   $\mathcal{G} \leftarrow \text{DFS}(KG, v_s, v_t)$;
3   $\mathcal{G}_{\text{set}}.\text{add}(\mathcal{G})$;
    // Randomly intervene on each node except v_t
4   **for** *each node v in $\mathcal{G}$* **do**
5   **if** $v \neq v_t$ **then**
6       $\widetilde{\mathcal{G}} \leftarrow \text{copy}(\mathcal{G})$;
7   Remove v and its edges in $\widetilde{\mathcal{G}}$;
8       $\mathcal{G}_{\text{set}}.\text{add}(\widetilde{\mathcal{G}})$;
9   **end**
10   **end**

11   **return** $\mathcal{G}_{\text{set}}$;

● **Factual Group:** Based on the principle of *Association*, we employ the depth-first search (DFS) algorithm to extract a benchmark factual graph $\mathcal{G}$ from a financial KG. $\mathcal{G}$ records the association with regard to risk contagion, including the causal parts X and the non-causal parts Z. Without any intervention, the probability of risk contagion is 100%.

● **Intervention Group:** Following the *Intervention* and *Counterfactual* rule, we introduce random do-operators within $\mathcal{G}$, following Algorithm 1. This approach entails directly manipulating specific variables within $\mathcal{G}$ to validate the effects of these interventions on risk propagation. For example, the probability of risk contagion is 0% after removing *Company B*. By comparing with the factual group, we observe a significant change in the probability of risk contagion before and after the intervention on *Company B*. Therefore, we can confirm that *Company B* is a contributing factor to the risk contagion. In this way, we can clearly distinguish between the causal and non-causal components.

In the explanation, we employ the chain-of-thought approach to guide LLMs in making causal inferences. As a result, we obtain a causal instruction set $\mathcal{I} = \{q_i, e_i\}_{i=1}^{N}$, where q and e represent the query and explanation, respectively.

3.3 Fusion Module

To integrate LLMs reasoning with KGs at the architectural level, this study introduces a fusion module f_θ, as depicted in Fig. 4. This module first ensures correct alignments between tokens and nodes via multi-scale contrastive learning. Subsequently, it achieves a deep integration of textual and graph-structured information through cross multi-head attention mechanisms.
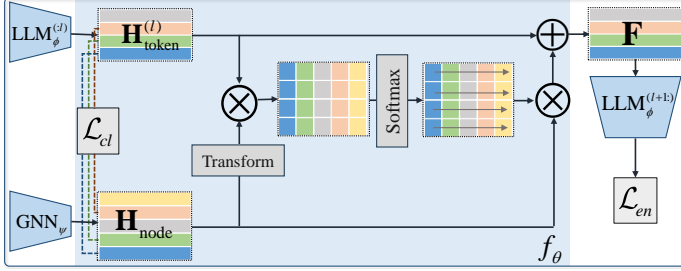

Fig. 4: Fusion Module f_θ

3.3.1 Multi-scale Contrastive Learning for Correct Alignments

Correct alignments between tokens and nodes guarantee that LLM_ϕ carries out accurate causal reasoning with KGs. Here, we use a multi-scale contrastive loss function to guide models in learning alignments at varying granularities.

$$\mathbf{H}_{\text{token}}^{(l)} = \text{LLM}_\phi^{(:l)}([q;e]) \in \mathbb{R}^{L \times d},$$
$$\mathbf{H}_{\text{node}} = \text{GNN}_\psi(\mathcal{G}_{\text{set}}) \in \mathbb{R}^{M \times d},$$

where $[q;e]$ denotes the concatenation of queries and explanations with a token delimiter, like $<\text{SEP}>$. $\text{LLM}_\phi^{(:l)}$ represents the output of the l-th layer of LLM_ϕ. L represents the number of tokens. M represents the number of nodes in $\mathcal{G}_{\text{set}}$.

$$\mathcal{L}_{\text{token\_node}} =$$
$$-\log \frac{\exp\left(\mathbb{1}_{[\text{paired}]}\text{sim}(\mathbf{H}_{\text{token}}^{(l)}[i], \mathbf{H}_{\text{node}}[i])/\tau\right)}{\sum_{i=1}^{\min(L,M)} \mathbb{1}_{[\text{unpaired}]}\exp\left(\text{sim}(\mathbf{H}_{\text{token}}^{(l)}[i], \mathbf{H}_{\text{node}}[i])/\tau\right)},$$

$$\mathcal{L}_{\text{token\_subgraph}} =$$
$$-\log \frac{\exp\left(\mathbb{1}_{[\text{paired}]}\text{sim}(\mathbf{H}_{\text{token}}^{(l)}[i], G_i)/\tau\right)}{\sum_{i=1}^{\min(L,M)} \mathbb{1}_{[\text{unpaired}]}\exp\left(\text{sim}(\mathbf{H}_{\text{token}}^{(l)}[i], G_i)/\tau\right)}.$$

In the above equation, $\mathcal{L}_{\text{token\_node}}$ denotes the contrastive loss between tokens and nodes. $\mathcal{L}_{\text{token\_subgraph}}$ captures the contrastive loss between tokens and k-hop subgraphs surrounding the nodes. $\mathbb{1}_{[\text{paired}]}$ serves as a predefined indicator, assigned a value of 1 if the i-th token and the i-th node are matched, and 0 otherwise. For example, if the token position and node position corresponding to *Company A* are the same, then $\mathbb{1}_{[\text{paired}]}$ is assigned a value of 1. $\mathbb{1}_{[\text{unpaired}]}$ represents the complementary indicator. $\mathbf{H}_{\text{token}}^{(l)}[i]$ and $\mathbf{H}_{\text{node}}[i]$ correspond to the embeddings of the i-th token and the i-th node, respectively. τ is the temperature parameter that modulates the scale of similarity scores between tokens and nodes, influencing the smoothness of the resulting probability distribution. A lower τ value emphasizes minor score discrepancies, leading to a more concentrated distribution, whereas a higher value yields a more uniform distribution. G_i represents the embedding of the k-hop subgraphs associated with the i-th node. By considering the w-span context around the i-th token, we construct $\mathcal{L}_{\text{context\_node}}$ and $\mathcal{L}_{\text{context\_subgraph}}$. The following equation is the multi-scale contrastive loss function that effectively captures relationships at varying granularities, enhancing the model's ability to discern and leverage hierarchical information between tokens and nodes.

$$\mathcal{L}_{cl} = \mathcal{L}_{\text{token\_node}} + \mathcal{L}_{\text{token\_subgraph}} \\ + \mathcal{L}_{\text{context\_node}} + \mathcal{L}_{\text{context\_subgraph}}. \quad (1)$$

3.3.2 Soft Prompt with Cross Multi-head Attention Mechanisms

Here, we employ a multi-head attention mechanism with J heads to learn the information fusion between tokens and nodes.

$$\mathbf{Q}_j = \text{Q-Linear}_j(\mathbf{H}_{\text{token}}^{(l)}) \in \mathbb{R}^{L \times \frac{d}{J}}, \quad (2)$$

$$\mathbf{K}_j = \text{K-Linear}_j(\mathbf{H}_{\text{node}}) \in \mathbb{R}^{M \times \frac{d}{J}}, \quad (3)$$

$$\mathbf{V}_j = \text{V-Linear}_j(\mathbf{H}_{\text{node}}) \in \mathbb{R}^{M \times \frac{d}{J}}, \quad (4)$$

$$\mathbf{A}_j = \text{Softmax}\left(\frac{\mathbf{Q}_j\mathbf{K}_j^\top}{\sqrt{d/J}}\right) \in \mathbb{R}^{L \times M}, \quad (5)$$

$$\mathbf{P}_j = \mathbf{A}_j\mathbf{V}_j \in \mathbb{R}^{L \times \frac{d}{J}}, \quad (6)$$

$$\mathbf{P} = [\mathbf{P}_1\|\mathbf{P}_2\|\cdots\|\mathbf{P}_J] \in \mathbb{R}^{L \times d}, \quad (7)$$

$$\mathbf{F} = [\mathbf{P} \oplus \mathbf{H}_{\text{token}}^{(l)}] \in \mathbb{R}^{2L \times d}, \quad (8)$$

$$\mathbf{T} = \text{LLM}_\phi^{(l+1:)}(\mathbf{F}) \in \mathbb{R}^{L \times C}, \quad (9)$$

$$\mathcal{L}_{en} = -\frac{1}{L}\sum_{\iota=1}^{L}\sum_{c=1}^{C}\mathbf{I}_{\iota,c}\log(\mathbf{T}_{\iota,c}[L+1:]). \quad (10)$$

In the equations presented, the linear transformation Q-Linear$_j(\cdot)$ for the j-th attention head is initially applied to the token representations from the l-th layer of LLM_ϕ, producing query matrices $\mathbf{Q}_j$. Simultaneously, node representations are transformed linearly to create key matrices $\mathbf{K}_j$ and value matrices $\mathbf{V}_j$. Following this, the attention weight matrix $\mathbf{A}_j$ is calculated from the product of $\mathbf{Q}_j$ and $\mathbf{K}_j$, with the application of the Softmax$(\cdot)$ function. This weight matrix, together with $\mathbf{V}_j$, generates the output $\mathbf{P}_j$ for each attention head. The outputs $\{\mathbf{P}_j\}_{j=1}^J$ are then concatenated to form the soft prompt $\mathbf{P}$. The concatenation of $\mathbf{P}$ with $\mathbf{H}_{\text{token}}^{(l)}$ along the sequence dimension results in $\mathbf{F}$. This combined output is subsequently processed through LLM_ϕ, leading to the token probabilities $\mathbf{T}$. Finally, we optimize our model using the cross-entropy loss function, where $\mathbf{I}$ represents the ground-truth token labels. $\mathbf{T}[L+1:]$ represents the predicted probabilities of the last L tokens. C represents the vocabulary size.

3.4 Risk Pathway Inference Module

To infer the paths of risk propagation, we further conduct the following risk pathway inference module for calculating risk scores for each node in the graph.

$$\widehat{P}(v_s \to v_t|do(\nu = 1 \; or \; 0)) =$$
$$\text{Sigmoid}\left(\text{Readout}(\text{S-Linear}([\mathbf{A}_1, \mathbf{A}_2, \cdots, \mathbf{A}_J]^\top))\right),$$

$$\mathcal{L}_X = \mathbb{E}_{\nu \in X}[\widehat{P}(v_s \to v_t|do(\nu = 1))$$
$$- \widehat{P}(v_s \to v_t|do(\nu = 0))],$$
$$\mathcal{L}_Z = \mathbb{E}_{\nu \in Z}[\widehat{P}(v_s \to v_t|do(\nu = 1))$$
$$- \widehat{P}(v_s \to v_t|do(\nu = 0))],$$
$$\mathcal{L}_{path} = \mathcal{L}_Z - \mathcal{L}_X.$$

In the above equations, $\widehat{P}(v_s \to v_t|do(\nu = 1 \; or \; 0))$ denotes the estimated probabilities of risk contagion from the initiating node v_s to the target node v_t. The S-Linear$(\cdot)$ function maps a high-dimensional matrix into a 1-D space. The Readout$(\cdot)$ function is responsible for aggregating node-level features into graph-level features by computing the average of the node features. The variables X and Z signify causal and non-causal nodes, respectively. The notation $do(\nu = 0)$ signifies an intervention applied to node ν, effectively setting its state to zero. Our objective is to achieve a state where $\widehat{P}$ equals 0 upon intervention at causal nodes, and where $\widehat{P}$ remains 1 when intervening at non-causal nodes.
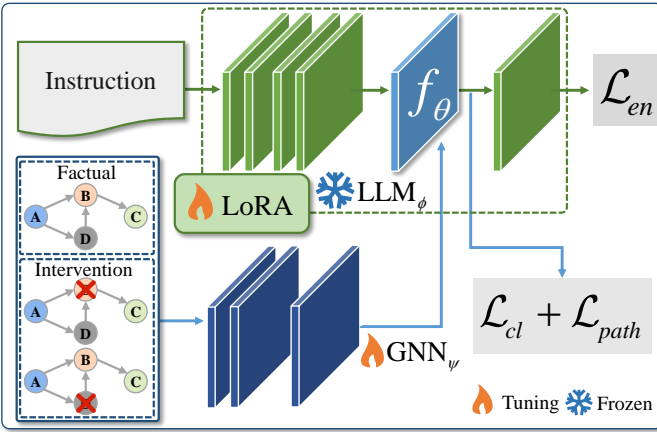
3.5 Joint Tuning

Fig. 5: Joint Fine-tuning of Our Proposed RC²R

To enable RC²R to effectively learn the causal mechanisms in financial risk propagation, we adopt the fine-tuning mechanism illustrated in Fig. 5. N pairs of instructions $\mathcal{I}$ and financial KGs $\mathcal{G}$ are input into LLM$_\phi$ and GNN$_\psi$, respectively. To train the overall model in a unified manner, we formulate the following joint loss function,

$$\mathcal{L}_{joint} = \mathcal{L}_{cl} + \mathcal{L}_{en} + \mathcal{L}_{path}. \tag{11}$$

Minimizing this loss function is designed to ensure precise alignments between textual and graph-based information, while also enabling LLMs to accurately generate explanations.

4 EXPERIMENTAL EVALUATION

In this section, we conduct a series of experiments on two open-sourced datasets to validate the performance of our proposed RC²R. We compare the effectiveness of predicted risk contagion and the quality of generated explanations. Additionally, we report and visualize the reasoning results of risk contagion pathways. Ultimately, we conduct ablation studies to verify the functionality and significance of each component within our proposed model.

4.1 Datasets

TABLE 2: Percentage of Financial Topics in Two Datasets

Datasets	Topics	Percentages
FinDKG	Stock	35%
	Bond	25%
	Money	20%
	Real estate	10%
	Commodity	10%
SupplyChain-KG	Raw materials	20%
	Manufacturing	30%
	Wholesale	20%
	Retail	30%

We open-source two comprehensive datasets, FinDKG and SupplyChain-KG, for the analysis of financial risk contagion. Each dataset encompasses causal texts and financial KGs.

• **FinDKG** [62]: This dataset covers 15 entity types and 15 relation types, encompassing a total of $13,645$ nodes and $242,149$ edges. Based on this open-source dataset, we have constructed $5,000$ causal instructions, covering diverse markets such as stock, bond, money, real estate, and commodity. The data ratios for each market are shown in Table 2. Following the guidelines in Table 1, we have designed tasks for each market, comprising 15% factual groups and 85% intervention groups.

• **SupplyChain-KG**: This dataset encompasses 10 entity types and 40 relation types, totaling $2,400$ nodes and $136,005$ edges. We have developed $5,000$ causal instructions addressing various risks within the supply chain, including financial shortages, natural and man-made disasters, and market volatility. Data distribution across different markets is detailed in Table 2. Following the guidelines in Table 1, we have crafted tasks for each supply chain sector, consisting of 17% factual groups and 83% intervention groups.

4.2 Experimental Settings

4.2.1 Baseline Models

Here, we provide the following four categories of baseline models, including LLMs, LLMs+KGs, GNNs, and GNN$_{cause}$.

• **LLMs**: **Gemma-7B** [63], which possesses a total parameter size of approximately 7 billion. The model is composed of 28 Transformer decoders, each with a hidden size of $3,072$ and 16 attention heads. Additionally, the feedforward

hidden dimension size for this model is $49,152$. The vocabulary size is set at $256,128$. **InternLM2-7B** [64], which possesses a total parameter size of approximately 7 billion, has demonstrated strong capabilities in text generation and comprehension. The model is composed of 32 Transformer decoder layers, each with a hidden size of $4,096$ and 32 attention heads. The vocabulary size is set $92,544$. **Llama2-7B** [65], is a pretrained generative text model with scales of 7 billion parameters. The model is composed of 32 Transformer blocks, each with a hidden size of $4,096$ and 32 attention heads. The vocabulary size is set at $32,000$. The carefully designed prompts for LLMs are reflected in Table 3.

• **LLMs+KGs**: **Gemma+**$t_{\mathcal{G}}$ is a combined framework of Gemma and KG information. In order to contextualize the large language model, we transform the KG information into triplet text $t_{\mathcal{G}}$ that serves as a context for model input. The elaborated prompts for LLMs+KGs are reflected in Table 3.

• **GNNs**: The Graph Convolutional Network (**GCN**) [66] applies convolutional operations on the nodes of a graph, enabling each node to aggregate information from its neighboring nodes, thereby learning the representation of each node. Graph Attention Network (**GAT**) [67] leverages attention mechanisms to weigh the importance of neighboring nodes in graph-structured data, enabling feature aggregation with varying emphasis on different neighbors.

• **GNN**$_{cause}$: The Discovering Invariant Rationale (**DIR**) method [39] aims to uncover causal rationales that remain consistent across diverse distributions, thereby facilitating the development of inherently interpretable GNNs. **V-REx** [40] represents a streamlined adaptation of risk extrapolation. It effectively minimizes risk disparities across training environments, and diminishes model vulnerability to extensive distributional shifts.

TABLE 3: Prompts Used for LLMs & LLMs+KGs

Role	You are a powerful causal reasoning assistant.
Triples $t_{\mathcal{G}}$	[Company A, partners with, Company B], ..., [Company E, supplies to, Company C]
Query	Company A produces silk, ..., Company C produces qipao. In Jan, Company A was at risk. In May, Company C was at risk. Please infer the risk contagion pathways from Company A to Company C.

* Notably, the LLMs+KGs baseline models utilize prompts that include role, triples, and query, whereas the LLMs baseline models employ prompts comprising only role and query.

4.2.2 Parameter Configuration of Our Model

According to our preliminary experiments, we choose the following parameters to ensure optimal performance. Our selection of LLMs includes the Gemma-7B [63], specifically fine-tuned via the LoRA method [68], to achieve enhanced model adaptability and performance. For GNNs, we prioritize models incorporating residual connections, notably the GCN [66]. The neural network architecture consists of 4 layers, a configuration that balances complexity with performance efficacy. We standardize the dimension of hidden layers at 1024, a decision guided by initial tests.

In multi-scale contrastive learning, we set $k = 2$ for subgraph extraction, $w = 3$ for context selection, and set τ to 1. Our configuration of the cross-attention mechanisms includes 8 heads. A batch size of 1 is selected, an unconventional choice that our empirical evidence suggests maximizes training efficiency and model convergence. The training spans 5 epochs, with a learning rate set at 0.01. Furthermore, to assess the effectiveness of cross attention mechanisms and multi-scale contrastive learning, we develop two variants: **RC²R**†, which replaces cross attention mechanisms with concatenation, and **RC²R**‡, which omits multi-scale contrastive learning.

4.2.3 Evaluation Metrics

To evaluate the effectiveness of predicted risk contagion, we measure them using the accuracy (**ACC**) and area under the curve (**AUC**) metrics.

To assess the quality of generated explanations, we consider the following five key criteria. (1) **Fluency** measures the naturalness and readability, determining if the text flows smoothly. (2) **Relevance** examines how well the text aligns with the given task or query. (3) **Consistency** checks for any contradictions within the information presented. (4) **Diversity** evaluates the range of different yet relevant outputs the model can produce. (5) **Informativeness** gauges the richness of useful information within the text. Additionally, we will report the average scores ($[0, 10]$) for each criterion based on both human evaluation and GPT-4 assessment, with higher scores indicating better performance.

To assess the predictive accuracy of risk propagation pathways, we employ the intersection over union (**IoU**) metric, defined as $\text{IoU} = \frac{\text{path intersection}}{\text{path union}} \in [0, 1]$. Here, the intersection denotes the count of nodes that overlap between the predicted and ground-truth paths, whereas the union represents the aggregate count of unique nodes obtained by combining those from both the predicted paths and the ground-truth paths. A higher IoU indicates better performance.

4.2.4 Experimental Platform

All experiments are conducted on a Linux server with a GPU (NVIDIA A800, Memory 80G) and CPU (Hygon C86 7375 32-core Processor). We implement our proposed model with deep learning library PyTorch, transformers, and Deep Graph Library (dgl). The versions of Python, PyTorch, transformers, and dgl are 3.8.10, 2.2.0+cu118, 4.38.1, 1.1.2+cu118, respectively.

4.3 Experimental Results & Analysis

In this section, we evaluate the performance of our proposed RC²R from three perspectives: predicted risk contagion, generated explanations, and inferred propagation pathways.

4.3.1 Performance of Predicted Risk Contagion

From Tables 4 and 5, we observe that RC²R outperforms Gemma+$t_{\mathcal{G}}$, achieving an average increase of 3.9% in ACC and 5.25% in AUC. This improvement is attributed to the ability of our model to effectively leverage the structural information in KGs via GNNs, whereas Gemma+$t_{\mathcal{G}}$ destroys

TABLE 4: ACC & AUC of Predicted Risk Contagion

	Dataset	FinDKG		SupplyChain-KG	
Category	Metric	ACC	AUC	ACC	AUC
LLMs+KGs	**RC²R (Our Model)**	**0.783** ± 0.037	**0.757** ± 0.045	**0.667** ± 0.021	**0.630** ± 0.030
	RC²R†	0.710 ± 0.028	0.671 ± 0.033	0.622 ± 0.024	0.610 ± 0.019
	RC²R‡	0.717 ± 0.041	0.697 ± 0.052	0.632 ± 0.029	0.619 ± 0.031
	Gemma+$t_\mathcal{G}$	0.747 ± 0.022	0.671 ± 0.044	0.625 ± 0.036	0.611 ± 0.027
LLMs	Gemma-7B [63]	0.667 ± 0.031	0.549 ± 0.039	0.603 ± 0.028	0.584 ± 0.024
	InternLM2-7B [64]	0.663 ± 0.025	0.608 ± 0.049	0.592 ± 0.037	0.621 ± 0.053
	Llama2-7B [65]	0.654 ± 0.023	0.620 ± 0.035	0.587 ± 0.031	0.546 ± 0.029
GNNs	GCN [66]	0.633 ± 0.027	0.579 ± 0.023	0.565 ± 0.036	0.524 ± 0.022
	GAT [67]	0.647 ± 0.041	0.596 ± 0.033	0.583 ± 0.029	0.523 ± 0.018
GNN$_{cause}$	DIR [39]	0.642 ± 0.038	0.571 ± 0.028	0.562 ± 0.025	0.521 ± 0.023
	V-REx [40]	0.672 ± 0.044	0.601 ± 0.039	0.573 ± 0.032	0.543 ± 0.031

TABLE 5: ACC & AUC of Predicted Risk Contagion on Out-of-distribution Data

	Training→Testing	SupplyChain-KG→FinDKG		FinDKG→SupplyChain-KG	
Category	Metric	ACC	AUC	ACC	AUC
LLMs+KGs	**RC²R (Our Model)**	**0.762** ± 0.021	**0.721** ± 0.055	**0.653** ± 0.019	**0.610** ± 0.034
	RC²R†	0.695 ± 0.014	0.654 ± 0.049	0.604 ± 0.026	0.591 ± 0.013
	RC²R‡	0.711 ± 0.017	0.677 ± 0.033	0.608 ± 0.027	0.581 ± 0.022
	Gemma+$t_\mathcal{G}$	0.725 ± 0.018	0.683 ± 0.042	0.612 ± 0.021	0.596 ± 0.013
LLMs	Gemma-7B [63]	0.611 ± 0.028	0.560 ± 0.045	0.600 ± 0.022	0.573 ± 0.019
	InternLM2-7B [64]	0.612 ± 0.035	0.605 ± 0.021	0.581 ± 0.044	0.548 ± 0.032
	Llama2-7B [65]	0.608 ± 0.039	0.580 ± 0.025	0.582 ± 0.033	0.541 ± 0.018
GNNs	GCN [66]	0.553 ± 0.048	0.531 ± 0.015	0.584 ± 0.042	0.554 ± 0.011
	GAT [67]	0.610 ± 0.031	0.547 ± 0.039	0.578 ± 0.047	0.553 ± 0.027
GNN$_{cause}$	DIR [39]	0.627 ± 0.036	0.568 ± 0.044	0.533 ± 0.015	0.518 ± 0.034
	V-REx [40]	0.619 ± 0.029	0.597 ± 0.022	0.558 ± 0.031	0.533 ± 0.026

the spatial structure of graphs by converting them into triplet texts. When compared to LLMs, RC²R demonstrates a significant average improvement of approximately 9.733% in ACC and 10.55% in AUC. This is largely due to the proficiency of our model in extracting crucial background information from KGs for causal reasoning. Against GNNs, our model shows a remarkable average improvement of 11.8% in ACC and 13.8% in AUC. This boost is a result of the successful integration of the semantic understanding and logical reasoning capabilities of LLMs. Furthermore, compared to GNN$_{cause}$, our model achieves an average improvement of 11.275% in ACC and 13.45% in AUC. This progress is due to the extensive causal instructions for model tuning, effectively guiding our model in conducting causal reasoning on risk propagation.

4.3.2 Quality of Generated Explanations

To assess the quality of explanations generated by RC²R, we combine human evaluation (by 8 evaluators) and GPT-4 review to compare our model with four baseline models. Each model is evaluated using 8 random samples. All evaluation results are ultimately aggregated into average values and presented in Tables 6 and 7. From them, we find that our model excels in several key indicators of explanation quality, especially in diversity, informativeness, and consistency. Specifically, our model shows significant improvements in terms of diversity, with an increase of 1.25. Similarly, it achieves a notable enhancement in informativeness, with an approximate increase of 0.28. Such advancements are

primarily attributed to the integration of knowledge content from the KGs, which enriches the explanatory depth of our model's output. Concurrently, there is an improvement in consistency, around 0.21, mainly due to contrastive learning facilitating effective alignment between text and nodes, ensuring that the content generated by the model remains consistent with the posed queries.
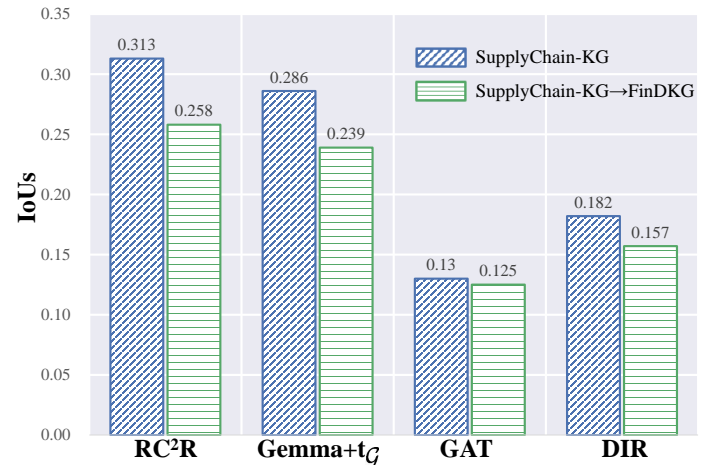
4.3.3 Performance of Inferred Propagation Pathways

Fig. 6: IoUs of Inferred Propagation Pathways

TABLE 6: Quality of Generated Explanations

Metrics	Diversity		Informativeness		Consistency		Relevance		Fluency	
Experts	Human	GPT-4	Human	GPT-4	Human	GPT-4	Human	GPT-4	Human	GPT-4
RC²R (Our Model)	**5.38** ± 0.60	**4.50** ± 0.20	**5.50** ± 0.55	6.25 ± 0.15	7.25 ± 0.45	8.50 ± 0.16	**8.00** ± 0.75	8.63 ± 0.15	8.13 ± 0.50	8.38 ± 0.05
Gemma+$t_{\mathcal{G}}$	5.13 ± 0.75	4.00 ± 0.20	5.13 ± 0.55	6.00 ± 0.20	7.00 ± 0.25	8.88 ± 0.05	8.00 ± 0.40	9.00 ± 0.13	8.63 ± 0.05	9.25 ± 0.10
Gemma-7B [63]	3.00 ± 0.55	2.63 ± 0.32	4.88 ± 0.24	6.13 ± 0.05	6.50 ± 0.25	8.50 ± 0.18	7.88 ± 0.47	8.50 ± 0.15	8.00 ± 0.50	9.00 ± 0.05
InternLM2-7B [64]	4.63 ± 0.45	4.00 ± 0.06	5.00 ± 0.55	7.00 ± 0.23	7.38 ± 0.60	8.00 ± 0.12	7.38 ± 0.48	8.50 ± 0.12	8.13 ± 0.62	9.00 ± 0.10
Llama2-7B [65]	3.00 ± 0.40	3.13 ± 0.20	4.50 ± 0.52	6.13 ± 0.11	6.88 ± 0.32	8.13 ± 0.08	7.38 ± 0.48	9.00 ± 0.05	8.00 ± 0.50	8.50 ± 0.15
Count	8		7		5		5		2	

* Count denotes the frequency at which our method outperforms other methods.

TABLE 7: Quality of Generated Explanations on Out-of-distribution Data (SupplyChain-KG→FinDKG)

Metrics	Diversity		Informativeness		Consistency		Relevance		Fluency	
Experts	Human	GPT-4	Human	GPT-4	Human	GPT-4	Human	GPT-4	Human	GPT-4
RC²R (Our Model)	**5.38** ± 0.32	4.13 ± 0.17	**5.13** ± 0.25	**6.00** ± 0.10	**7.25** ± 0.60	7.50 ± 0.22	7.63 ± 0.60	7.50 ± 0.08	8.38 ± 0.48	8.50 ± 0.18
Gemma+$t_{\mathcal{G}}$	4.50 ± 0.25	3.88 ± 0.04	4.50 ± 0.15	5.50 ± 0.05	6.63 ± 0.32	7.25 ± 0.12	8.13 ± 0.46	9.13 ± 0.19	8.50 ± 0.52	9.50 ± 0.20
Gemma-7B [63]	3.00 ± 0.20	3.00 ± 0.02	3.88 ± 0.15	3.13 ± 0.12	6.00 ± 0.22	7.25 ± 0.15	7.50 ± 0.45	8.50 ± 0.20	8.38 ± 0.40	9.00 ± 0.09
InternLM2-7B [64]	5.00 ± 0.22	4.38 ± 0.04	4.88 ± 0.10	5.63 ± 0.14	6.50 ± 0.40	8.00 ± 0.13	7.13 ± 0.32	8.00 ± 0.05	8.13 ± 0.33	8.88 ± 0.16
Llama2-7B [65]	3.50 ± 0.13	3.00 ± 0.07	4.50 ± 0.15	5.63 ± 0.03	6.50 ± 0.22	7.00 ± 0.10	6.88 ± 0.33	7.13 ± 0.03	8.13 ± 0.45	8.50 ± 0.20
Count	7		8		7		4		2	

In Fig. 6, RC²R outperforms Gemma+$t_{\mathcal{G}}$ with a 2.3% increase in the IoU score. This gain is linked to our model's effective integration of graph information, enhancing its ability to accurately identify pathways for inference propagation. Compared to GAT, our model secures a 15.8% improvement in the IoU score. This significant boost stems from the model's skillful leverage of inferential capabilities inherent in LLMs, moving beyond the exclusive reliance on the strengths of GNNs. When measured against DIR, our model shows an 11.6% increase in the IoU score. This enhancement primarily stems from our model's ability to parse textual semantics and comprehensively grasp various factual and intervention groups within causal instructions.

4.3.4 Visualization of Propagation Pathways

As illustrated in Fig. 7, our model provides detailed explanations when answering queries and visualizes the intensity and direction of financial risk contagion. The diagram highlights the path of risk contagion through orange nodes, specifically GlowShop (store)→ IllumiStore Retailers (retailer)→ RadiantShop (e-commerce platform) → Light-Fab (factory), which is the actual risk contagion path in the SupplyChain-KG dataset. The thickness of the path indicates the intensity of the risk contagion. For more cases, please refer to Figs. 8 and 9.

4.4 Ablation Study

4.4.1 Performance of Combining LLMs with KGs

From Table 4, we observe significant enhancements in ACC for both datasets through the LLMs+KGs approach, with increases of 11.1% and 6.529%, respectively. In Table 5, the LLMs+KGs methodology showcases substantial improvements in managing OOD data over the use of LLMs alone. Specifically, it achieved ACC performance boosts of 13.779% and 5.879% for the FinDKG and SupplyChain-KG datasets, respectively, along with AUC performance enhancements of 12.033% and 4.9%. These results underscore the efficacy of integrating KGs in boosting prediction accuracy, particularly in contexts requiring nuanced domain knowledge. This enhancement is likely attributed to the structured knowledge from KGs, which supports our model in achieving a deeper understanding and more effective reasoning.

In Table 4, the LLMs+KGs approaches demonstrate a significant performance enhancement over the GNNs methods, with ACC improvements of 12.5% and 7.2% for the FinDKG and SupplyChain-KG datasets, respectively, alongside AUC gains of 12.65% and 9.7%. From Table 5, it is evident that the LLMs+KGs approaches markedly surpass the GNNs methods in addressing OOD data, showcasing ACC and AUC enhancements of 16.2% and 16.3% for the FinDKG dataset, and 5.15% and 4.95% for the SupplyChain-KG dataset, respectively. These significant improvements are primarily attributable to the reasoning and generalization capabilities of LLMs.

4.4.2 Performance of Cross Attention Mechanisms

In Table 4, we observe that removing multi-head attention mechanisms results in decreased ACC and AUC metrics. Specifically, for the FinDKG dataset, the reductions in ACC and AUC are approximately 7.3% and 8.6%, respectively; for the SupplyChain-KG dataset, the decreases are around 4.5% and 2%, respectively. These findings suggest that multi-head attention mechanisms significantly affect the integration of KGs with LLMs, thereby impacting model performance.

4.4.3 Performance of Multi-scale Contrastive Loss

In Table 4, compared to RC²R‡, RC²R achieves approximately 6.6%, 6%, and 3.5%, 1.1% improvements in ACC and AUC performance on the FinDKG and SupplyChain-KG datasets, respectively. This significant leap in performance can largely be attributed to the multi-scale contrastive learning function, which effectively aligns fine-grained information between text and graphs.
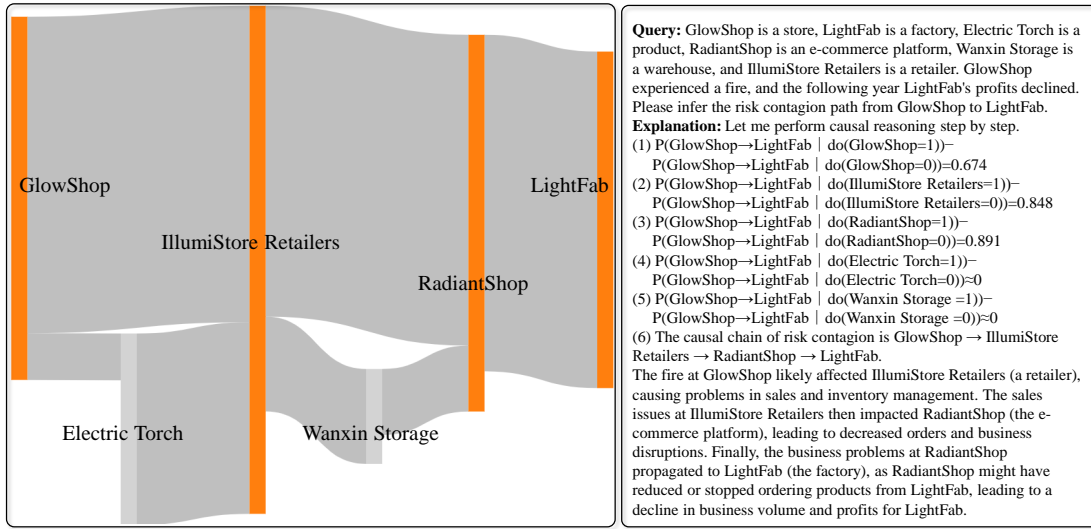
Query: GlowShop is a store, LightFab is a factory, Electric Torch is a product, RadiantShop is an e-commerce platform, Wanxin Storage is a warehouse, and IllumiStore Retailers is a retailer. GlowShop experienced a fire, and the following year LightFab's profits declined. Please infer the risk contagion path from GlowShop to LightFab.

Explanation: Let me perform causal reasoning step by step.
(1) P(GlowShop→LightFab | do(GlowShop=1))−
P(GlowShop→LightFab | do(GlowShop=0))=0.674
(2) P(GlowShop→LightFab | do(IllumiStore Retailers=1))−
P(GlowShop→LightFab | do(IllumiStore Retailers=0))=0.848
(3) P(GlowShop→LightFab | do(RadiantShop=1))−
P(GlowShop→LightFab | do(RadiantShop=0))=0.891
(4) P(GlowShop→LightFab | do(Electric Torch=1))−
P(GlowShop→LightFab | do(Electric Torch=0))≈0
(5) P(GlowShop→LightFab | do(Wanxin Storage =1))−
P(GlowShop→LightFab | do(Wanxin Storage =0))≈0
(6) The causal chain of risk contagion is GlowShop → IllumiStore Retailers → RadiantShop → LightFab.
The fire at GlowShop likely affected IllumiStore Retailers (a retailer), causing problems in sales and inventory management. The sales issues at IllumiStore Retailers then impacted RadiantShop (the e-commerce platform), leading to decreased orders and business disruptions. Finally, the business problems at RadiantShop propagated to LightFab (the factory), as RadiantShop might have reduced or stopped ordering products from LightFab, leading to a decline in business volume and profits for LightFab.

Fig. 7: Visualization of Risk Intensity via a Sankey Diagram & Causal Explanations: Case 1

Query: Smart-Controlled Doll is a product, Aurora Attic is a store, Toyland Wholesalers is a supplier, Leaf Wharf is a shipping company, Lili Toy Store is a store, and Blissful Bungalow is an online store. In January, Smart-Controlled Doll experienced product issues. In August, Leaf Wharf faced a financial shortfall. Please infer the risk contagion path from Smart-Controlled Doll to Leaf Wharf.

Explanation: Let me perform causal reasoning step by step.
(1) P(Smart-Controlled Doll → Leaf Wharf | do(Smart-Controlled Doll = 1))-
P(Smart-Controlled Doll → Leaf Wharf | do(Smart-Controlled Doll = 0))=0.724
(2) P(Smart-Controlled Doll → Leaf Wharf | do(Aurora Attic = 1))-
P(Smart-Controlled Doll → Leaf Wharf | do(Aurora Attic = 0))=0.883
(3) P(Smart-Controlled Doll → Leaf Wharf | do(Toyland Wholesalers = 1))-
P(Smart-Controlled Doll → Leaf Wharf | do(Toyland Wholesalers = 0))=0.922
(4) P(Smart-Controlled Doll → Leaf Wharf | do(Lili Toy Store = 1))-
P(Smart-Controlled Doll → Leaf Wharf | do(Lili Toy Store = 0)) ≈ 0
(5) P(Smart-Controlled Doll → Leaf Wharf | do(Blissful Bungalow = 1))-
P(Smart-Controlled Doll → Leaf Wharf | do(Blissful Bungalow = 0)) ≈ 0
(6) The causal chains of risk propagation is Smart-Controlled Doll → Aurora Attic → Toyland Wholesalersc → Leaf Wharf.
The product issue with Smart-Controlled Doll in January likely affected Aurora Attic (a store), causing problems in sales and inventory management. Aurora Attic's sales issues then impacted Toyland Wholesalers (the supplier), resulting in decreased orders and business disruptions. Finally, the business problems at Toyland Wholesalers propagated to Leaf Wharf (the shipping company), as Toyland Wholesalers might have reduced or stopped using Leaf Wharf's shipping services, leading to a decline in business volume and financial shortfall for Leaf Wharf.

Fig. 8: Visualization of Risk Intensity via a Sankey Diagram & Causal Explanations: Case 2

Query: OutFab and Bicycle Factory are manufacturers, TrailBulk is a wholesaler, Bicycle and Tent are products, CampLog is a logistics company, TrekStore is an online store, and GearMart is a retailer. In March, OutFab ceased production. The following year, TrekStore's popularity declined. Please infer the risk transmission path from OutFab to TrekStore.

Explanation: Let me perform causal reasoning step by step.
(1) P(OutFab→TrekStore | do(OutFab =1)) -
P(OutFab→TrekStore | do(OutFab =0)) = 0.771
(2) P(OutFab→TrekStore | do(Bicycle Factory =1)) -
P(OutFab→TrekStore | do(Bicycle Factory =0)) = 0.952
(3) P(OutFab→TrekStore | do(CampLog =1)) -
P(OutFab→TrekStore | do(CampLog =0)) = 0.913
(4) P(OutFab→TrekStore | do(Bicycle =1)) -
P(OutFab→TrekStore | do(Bicycle =0)) ≈ 0
(5) P(OutFab→TrekStore | do(TrailBulk =1)) -
P(OutFab→TrekStore | do(TrailBulk =0)) ≈ 0
(6) P(OutFab→TrekStore | do(Tent =1)) -
P(OutFab→TrekStore | do(Tent =0)) ≈ 0
(7) P(OutFab→TrekStore | do(GearMart =1)) -
P(OutFab→TrekStore | do(GearMart =0)) ≈ 0
(8) The causal chain of risk contagion is OutFab → Bicycle Factory → CampLog → TrekStore.
OutFab's production halt significantly impacted Bicycle Factory, which then affected CampLog, leading to a decline in TrekStore's popularity. This causal chain shows that OutFab's cessation of production impacted the manufacturing and logistics stages, ultimately affecting the online store.
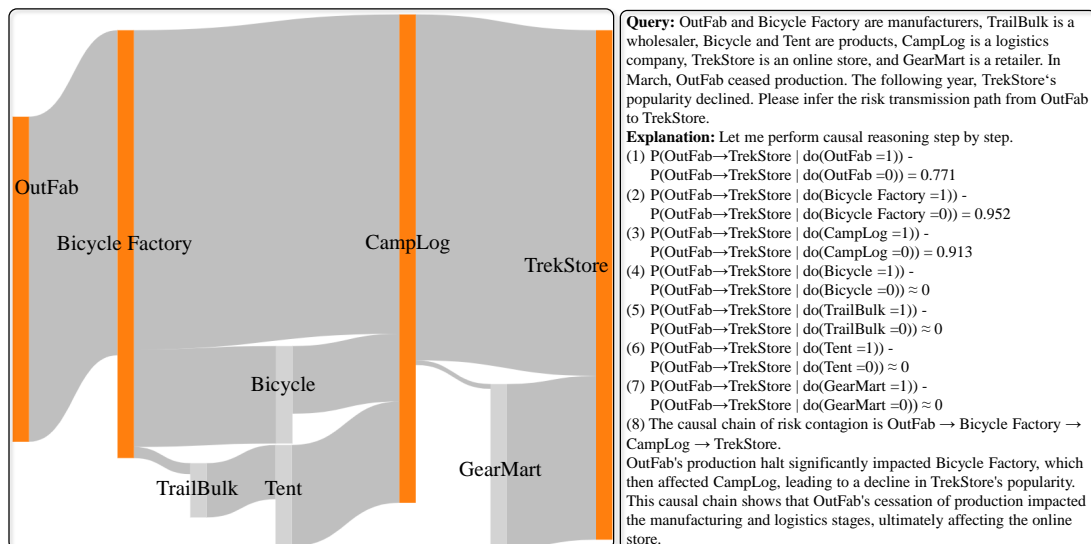
Fig. 9: Visualization of Risk Intensity via a Sankey Diagram & Causal Explanations: Case 3

4.4.4 Robustness Analysis of Causal Reasoning

To evaluate the robustness of our model's causal reasoning against the risk of contagion, we perform two tests on FinDKG: the *Random Confounder* and the *Subset of Data* tests [69]. In the *Random Confounder* test, we introduce additional confounding nodes into the factual graph to simulate potential external influences. The *Subset of Data* test involves the random removal of nodes from the factual graph, which helps us understand how our model copes with incomplete data sets.

As shown in Fig. 10, our model maintains consistent performance in the *Random Confounder* and *Subset of Data* tests compared to the *Estimated Effect* test, whereas the other three models exhibit significant declines under the same conditions. This underscores the robust causal reasoning of our model. Such robustness primarily stems from the diverse interventions we implemented and the explicit incorporation of influence estimation for fine-grained causal features in $\mathcal{L}_{path}$. In contrast, Gemma+t_g and DIR lack these design mechanisms. Additionally, the attention mechanism employed by GAT is based on correlations that lack the robustness of causal inference.
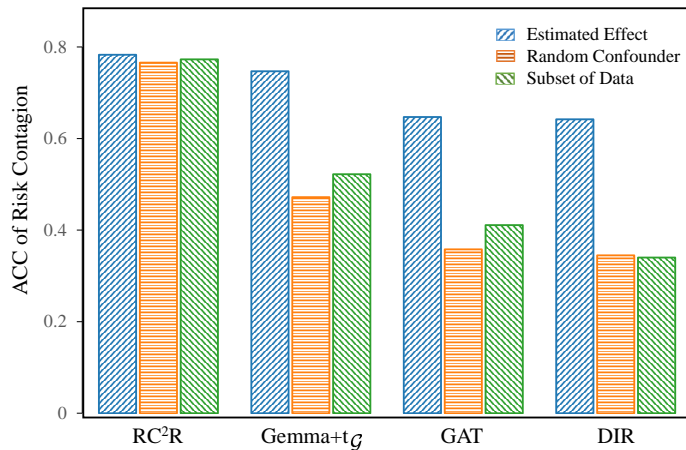

Fig. 10: Robustness Test of Causal Reasoning on FinDKG

5 Conclusion

To analyze the causal effects behind risk contagion, this study proposes RC^{2}R, which combines the reasoning capabilities of LLMs with the factual and professional knowledge within KGs, aiming to infer the causal relationships in financial risk contagion. At the data level, we rigorously follow the hierarchies of causation to develop causal instructions for fine-tuning our model and activating the causal reasoning capabilities of LLMs. In terms of model architecture level, we develop a fusion module integrating textual and graph information through a multi-scale contrastive loss function and soft prompt with multi-head attention mechanisms. Furthermore, we establish a risk pathway inference module to identify propagation routes, which are visualized via Sankey diagrams. In the near future, we intend to develop agents grounded in LLMs to autonomously conduct formal causal reasoning. We aim to apply this technology to analyze the causal factors behind financial risk contagion [70], [71].

References

[1] P. Gai and S. Kapadia, "Contagion in financial networks," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 466, no. 2120, pp. 2401–2423, 2010.

[2] M. Elliott, B. Golub, and M. O. Jackson, "Financial networks and contagion," *American Economic Review*, vol. 104, no. 10, pp. 3115–3153, 2014.

[3] L. Eisenberg and T. H. Noe, "Systemic risk in financial systems," *Management Science*, vol. 47, no. 2, pp. 236–249, 2001.

[4] M. Billio, M. Getmansky, A. W. Lo, and L. Pelizzon, "Econometric measures of connectedness and systemic risk in the finance and insurance sectors," *Journal of Financial Economics*, vol. 104, no. 3, pp. 535–559, 2012.

[5] A. Gepp and K. Kumar, "Predicting financial distress: A comparison of survival analysis and decision tree techniques," *Procedia Computer Science*, vol. 54, pp. 396–404, 2015.

[6] M. Malekipirbazari and V. Aksakalli, "Risk assessment in social lending via random forests," *Expert Systems with Applications*, vol. 42, no. 10, pp. 4621–4631, 2015.

[7] P. Danenas and G. Garsva, "Selection of support vector machines based classifiers for credit risk domain," *Expert systems with applications*, vol. 42, no. 6, pp. 3194–3204, 2015.

[8] Y.-C. Chang, K.-H. Chang, H.-H. Chu, and L.-I. Tong, "Establishing decision tree-based short-term default credit risk assessment models," *Communications in Statistics-Theory and Methods*, vol. 45, no. 23, pp. 6803–6815, 2016.

[9] H. Jiang, W.-K. Ching, K. F. C. Yiu, and Y. Qiu, "Stationary Mahalanobis kernel SVM for credit risk evaluation," *Applied Soft Computing*, vol. 71, pp. 407–417, 2018.

[10] N. Arora and P. D. Kaur, "A bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment," *Applied Soft Computing*, vol. 86, p. 105936, 2020.

[11] T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," *European journal of operational research*, vol. 270, no. 2, pp. 654–669, 2018.

[12] T. Hosaka, "Bankruptcy prediction using imaged financial ratios and convolutional neural networks," *Expert systems with applications*, vol. 117, pp. 287–299, 2019.

[13] D. Wang, J. Lin, P. Cui, Q. Jia, Z. Wang, Y. Fang, Q. Yu, J. Zhou, S. Yang, and Y. Qi, "A semi-supervised graph attentive network for financial fraud detection," in *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2019, pp. 598–607.

[14] Q. Li, J. Tan, J. Wang, and H. Chen, "A multimodal event-driven LSTM model for stock prediction using online news," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 10, pp. 3323–3337, 2020.

[15] B. Xu, H. Shen, B. Sun, R. An, Q. Cao, and X. Cheng, "Towards consumer loan fraud detection: Graph neural networks with role-constrained conditional random field," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 5, 2021, pp. 4537–4545.

[16] Z.-s. Ouyang, Y. Lai *et al.*, "Systemic financial risk early warning of financial market in china using Attention-LSTM model," *The North American Journal of Economics and Finance*, vol. 56, p. 101383, 2021.

[17] D. Cheng, Z. Niu, J. Li, and C. Jiang, "Regulating systemic crises: Stemming the contagion risk in networked-loans through deep graph learning," *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[18] D. Cheng, F. Yang, S. Xiang, and J. Liu, "Financial time series forecasting with multi-modality graph neural network," *Pattern Recognition*, vol. 121, p. 108218, 2022.

[19] C. Zhang, S. Bauer, P. Bennett, J. Gao, W. Gong, A. Hilmkil, J. Jennings, C. Ma, T. Minka, N. Pawlowski *et al.*, "Understanding causality with large language models: Feasibility and opportunities," *arXiv preprint arXiv:2304.05524*, 2023.

[20] T. Webb, K. J. Holyoak, and H. Lu, "Emergent analogical reasoning in large language models," *Nature Human Behaviour*, vol. 7, no. 9, pp. 1526–1541, 2023.

[21] L. Yang, H. Chen, Z. Li, X. Ding, and X. Wu, "Give us the facts: Enhancing large language models with knowledge graphs for fact-aware language modeling," *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[22] H. Chen, B. Liao, J. Luo, W. Zhu, and X. Yang, "Learning a structural causal model for intuition reasoning in conversation," *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[23] M. Loster, T. Repke, R. Krestel, F. Naumann, J. Ehmueller, B. Feldmann, and O. Maspfuhl, "The challenges of creating, maintaining and exploring graphs of financial entities," in *Proceedings of the Fourth International Workshop on Data Science for Macro-Modeling with Financial and Economic Datasets*, 2018, pp. 1–2.

[24] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, "Unifying large language models and knowledge graphs: A roadmap," *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[25] W. Lu, J. Li, Y. Li, A. Sun, and J. Wang, "A CNN-LSTM-based model to forecast stock prices," *Complexity*, vol. 2020, no. 1, p. 6622927, 2020.

[26] Y.-C. Chen and W.-C. Huang, "Constructing a stock-price forecast CNN model with gold and crude oil indicators," *Applied Soft Computing*, vol. 112, p. 107760, 2021.

[27] S. Cavalli and M. Amoretti, "CNN-based multivariate data analysis for bitcoin trend prediction," *Applied Soft Computing*, vol. 101, p. 107065, 2021.

[28] H. Wang, J. Wang, L. Cao, Y. Li, Q. Sun, and J. Wang, "A stock closing price prediction model based on CNN-BiSLSTM," *Complexity*, vol. 2021, no. 1, p. 5360828, 2021.

[29] D. Cheng, Z. Niu, and Y. Zhang, "Contagious chain risk rating for networked-guarantee loans," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 2715–2723.

[30] Y. Rao, X. Mi, C. Duan, X. Ren, J. Cheng, Y. Chen, H. You, Q. Gao, Z. Zeng, and X. Wei, "Know-gnn: An explainable knowledge-guided graph neural network for fraud detection," in *International Conference on Neural Information Processing*. Springer, 2021, pp. 159–167.

[31] M. Duan, T. Zheng, Y. Gao, G. Wang, Z. Feng, and X. Wang, "DGA-GNN: Dynamic grouping aggregation GNN for fraud detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 10, 2024, pp. 11 820–11 828.

[32] L. Yang, T. L. J. Ng, B. Smyth, and R. Dong, "Html: Hierarchical transformer-based multi-task learning for volatility prediction," in *Proceedings of The Web Conference 2020*, 2020, pp. 441–451.

[33] L. Yang, J. Li, R. Dong, Y. Zhang, and B. Smyth, "Numhtml: Numeric-oriented hierarchical transformer model for multi-task financial forecasting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 11 604–11 612.

[34] A. H. Huang, H. Wang, and Y. Yang, "FinBERT: A large language model for extracting information from financial text," *Contemporary Accounting Research*, vol. 40, no. 2, pp. 806–841, 2023.

[35] M. Leippold, "Sentiment spin: Attacking financial sentiment with gpt-3," *Finance Research Letters*, p. 103957, 2023.

[36] J. Brehmer, P. De Haan, P. Lippe, and T. S. Cohen, "Weakly supervised causal representation learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 38 319–38 331, 2022.

[37] W. Wang, X. Lin, F. Feng, X. He, M. Lin, and T.-S. Chua, "Causal representation learning for out-of-distribution recommendation," in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 3562–3571.

[38] N. Pawlowski, D. Coelho de Castro, and B. Glocker, "Deep structural causal models for tractable counterfactual inference," *Advances in neural information processing systems*, vol. 33, pp. 857–869, 2020.

[39] Y. Wu, X. Wang, A. Zhang, X. He, and T.-S. Chua, "Discovering invariant rationales for graph neural networks," in *International Conference on Learning Representations*, 2021.

[40] D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. Le Priol, and A. Courville, "Out-of-distribution generalization via risk extrapolation," in *International Conference on Machine Learning*. Pmlr, 2021, pp. 5815–5826.

[41] N. Naik, A. Khandelwal, M. Joshi, M. Atre, H. Wright, K. Kannan, S. Hill, G. Mamidipudi, G. Srinivasa, C. Bifulco *et al.*, "Applying large language models for causal structure learning in non small cell lung cancer," *arXiv preprint arXiv:2311.07191*, 2023.

[42] E. Kıcıman, R. Ness, A. Sharma, and C. Tan, "Causal reasoning and large language models: Opening a new frontier for causality," *arXiv preprint arXiv:2305.00050*, 2023.

[43] A. Feder, Y. Wald, C. Shi, S. Saria, and D. Blei, "Causal-structure driven augmentations for text ood generalization," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[44] X. Liu, P. Xu, J. Wu, J. Yuan, Y. Yang, Y. Zhou, F. Liu, T. Guan, H. Wang, T. Yu *et al.*, "Large language models and causal inference in collaboration: A comprehensive survey," *arXiv preprint arXiv:2403.09606*, 2024.

[45] S. Long, T. Schuster, A. Piché, S. Research *et al.*, "Can large language models build causal graphs?" *arXiv preprint arXiv:2303.05279*, 2023.

[46] R. Tu, C. Ma, and C. Zhang, "Causal-discovery performance of chatgpt in the context of neuropathic pain diagnosis," *arXiv preprint arXiv:2301.13819*, 2023.

[47] Z. Chen, Q. Gao, A. Bosselut, A. Sabharwal, and K. Richardson, "Disco: Distilling counterfactuals with large language models," *arXiv preprint arXiv:2212.10534*, 2022.

[48] J. Zhang, J. Jennings, C. Zhang, and C. Ma, "Towards causal foundation model: on duality between causal inference and attention," *arXiv preprint arXiv:2310.00809*, 2023.

[49] A. Gunjal, J. Yin, and E. Bas, "Detecting and preventing hallucinations in large vision language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 16, 2024, pp. 18 135–18 143.

[50] X. Guan, Y. Liu, H. Lin, Y. Lu, B. He, X. Han, and L. Sun, "Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 16, 2024, pp. 18 126–18 134.

[51] A. Zeng, M. Liu, R. Lu, B. Wang, X. Liu, Y. Dong, and J. Tang, "Agenttuning: Enabling generalized agent abilities for llms," *arXiv preprint arXiv:2310.12823*, 2023.

[52] J. Jiang, K. Zhou, Z. Dong, K. Ye, W. X. Zhao, and J.-R. Wen, "Structgpt: A general framework for large language model to reason over structured data," *arXiv preprint arXiv:2305.09645*, 2023.

[53] J. Sun, C. Xu, L. Tang, S. Wang, C. Lin, Y. Gong, H.-Y. Shum, and J. Guo, "Think-on-graph: Deep and responsible reasoning of large language model with knowledge graph," *arXiv preprint arXiv:2307.07697*, 2023.

[54] B. Y. Lin, X. Chen, J. Chen, and X. Ren, "Kagnet: Knowledge-aware graph networks for commonsense reasoning," *arXiv preprint arXiv:1909.02151*, 2019.

[55] M. Yasunaga, H. Ren, A. Bosselut, P. Liang, and J. Leskovec, "QA-GNN: Reasoning with language models and knowledge graphs for question answering," *arXiv preprint arXiv:2104.06378*, 2021.

[56] Y. Sun, Q. Shi, L. Qi, and Y. Zhang, "JointLK: Joint reasoning with language models and knowledge graphs for commonsense question answering," *arXiv preprint arXiv:2112.02732*, 2021.

[57] M. Zečević, M. Willig, D. S. Dhami, and K. Kersting, "Causal parrots: Large language models may talk causality but are not causal," *Transactions on Machine Learning Research*, 2023.

[58] Z. Jin, Y. Chen, F. Leeb, L. Gresele, O. Kamal, Z. Lyu, K. Blin, F. Gonzalez Adauto, M. Kleiman-Weiner, M. Sachan *et al.*, "Cladder: A benchmark to assess causal reasoning capabilities of language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[59] J. Pearl, M. Glymour, and N. P. Jewell, *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[60] J. Pearl and D. Mackenzie, *The book of why: The new science of cause and effect*. Basic Books, 2018.

[61] M. Goldszmidt and J. Pearl, "Rank-based systems: A simple approach to belief revision, belief update, and reasoning about evidence and actions." *KR*, vol. 92, pp. 661–672, 1992.

[62] X. V. Li, "FinDKG: Dynamic knowledge graph with large language models for global finance," *Available at SSRN 4608445*, 2023.

[63] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, L. Sifre, M. Rivière, M. S. Kale, J. Love, P. Tafti, and et al., "Gemma," 2024. [Online]. Available: https://www.kaggle.com/m/3301

[64] I. Team, "Internlm: A multilingual language model with progressively enhanced capabilities," 2023.

[65] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv*, pp. arXiv–2307, 2023.

[66] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations*, 2016.

[67] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations*, 2018.

[68] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2021.

[69] E. Kiciman and A. Sharma, "Tutorial on causal inference and counterfactual reasoning," in *ACM KDD International Conference on Knowledge Discovery and Data Mining*, 2018.

[70] B. Algieri and A. Leccadito, "Assessing contagion risk from energy and non-energy commodity markets," *Energy Economics*, vol. 62, pp. 312–322, 2017.

[71] X. Gong and J. Xu, "Geopolitical risk and dynamic connectedness between commodity markets," *Energy Economics*, vol. 110, p. 106028, 2022.