SURF: A System to Unveil Explainable Risk Relations between Firms

Yu-Hsiang Wang[1*], Wei-Ning Chiu[1,2*], Yi-Tai Hsiao[1], Yu-Shiang Huang[2], Yi-Shyuan Chiang[3], Shuo-En Wu[1], Chuan-Ju Wang[1]

[1]Academia Sinica, [2]National Taiwan University, [3]University of Illinois at Urbana-Champaign

Correspondence: cjwang@citi.sinica.edu.tw

Abstract

Firm risk relations are crucial in financial applications, including hedging and portfolio construction. However, the complexity of extracting relevant information from financial reports poses significant challenges in quantifying these relations. To this end, we introduce SURF, a **S**ystem to **U**nveil Explainable Risk **R**elations between **F**irms. SURF employs a domain-specific encoder and an innovative scoring mechanism to uncover latent risk connections from financial reports. It constructs a network graph to visualize these firm-level risk interactions and incorporates a rationale explainer to elucidate the underlying links. Our evaluation using stock data shows that SURF outperforms baseline methods in effectively capturing firm risk relations. The demo video of the system is publicly available.[1]

1 Introduction and Related Work

Financial markets are shaped by many factors, including market conditions and regulatory policies. The intricacies of these influences necessitate expert analysis to assess their impacts accurately. A critical component of such analysis is understanding the strength of risk relations between firms, a key to supporting investors and financial professionals in making well-informed decisions. Risk relations refer to the connections between two companies based on shared risk exposure. Marriott and Hilton, for example, have strong risk relations because they face similar risks, such as competition from other hotel operators and the occurrence of disasters. Traditionally, financial analysts manually review extensive financial reports to identify these relations. However, this approach is time-intensive and prone to subjective bias, making efficiently conducting large-scale analyses significantly challenging.

Previous studies have used deep neural networks to classify relations or predict binary links between firms (Wichmann et al., 2020; Kosasih and Brintrup, 2022). However, these methods often lack interpretability or fail to quantify the strength of relations. In the field of finance, explainability is crucial to users' trust, highlighting the need for systematic approaches that improve transparency and demystify the "black box" nature of these models (Bracke et al., 2019; Hoepner et al., 2021).

To measure and explain the strength of risk relations between firms, we present an interactive system that extracts information from key risk-related sections of Form 10-K filings—specifically "Item 1. Business," "Item 1A. Risk Factors," and "Item 7A. Quantitative and Qualitative Disclosures about Market Risk"—to provide comprehensive insights into inter-firm risk relations.[2] These sections typically use lexically similar terms, cover related topics, and often detail risk events along with their dates. Building on these characteristics, we design a custom encoder and implement a novel retrieval-based approach to identify chronological and lexical similarities among the paragraphs. We define "mutual risk paragraphs (abbreviated as MRPs, hereafter)" as those discussing similar risks and compute a risk relation score between two firms based on the proportion of MRPs they share. Our system computes and explains these scores by summarizing the identified MRPs using large language models (LLMs), enabling users to interactively explore and intuitively understand insights about shared risks between firms. By simultaneously presenting the LLM-generated summaries alongside factual paragraphs from 10-K filings, SURF further mitigates the risks of hallucination and factual inconsistencies. Meanwhile, user feedback is collected and incorporated to dynamically enhance

[*]These authors contributed equally to this work.

[1]https://www.youtube.com/watch?v=HobCyNgR9T0

[2]Form 10-K filings are annual reports required by the U.S. Securities and Exchange Commission (SEC).

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations), pages 260–267

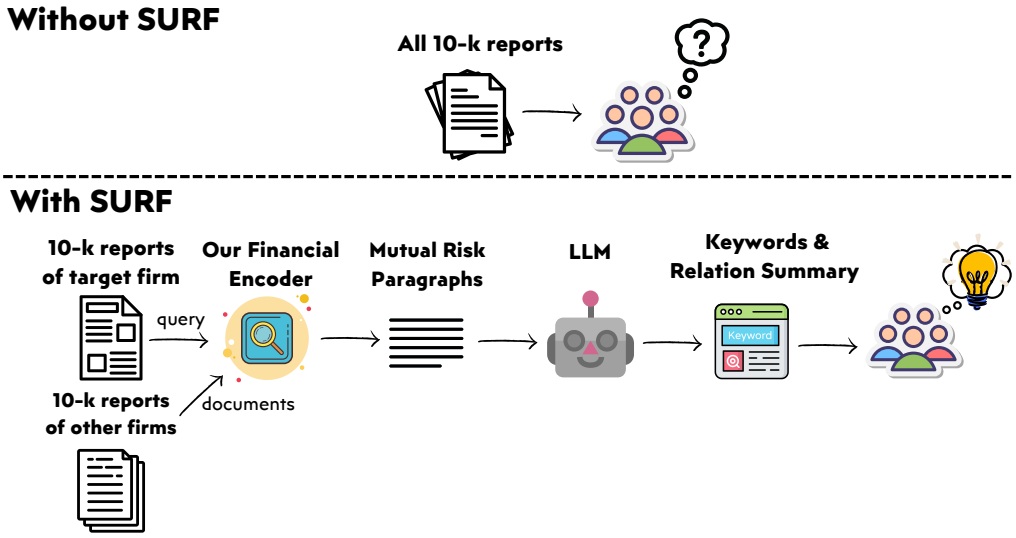

Figure 1: A conceptual overview of the contribution of SURF

the rationale explainer. The system's ability to reveal latent relations is further validated through experiments and a proposed metric. A conceptual overview of SURF's contribution is presented in Figure 1.

2 Core Functionalities of SURF

SURF[3] is an interactive platform designed to visualize risk relations between firms and explain the rationale behind the connections. It features three main components, as shown in Figure 2.

User Setting Module This module (labeled as partition (1) in Figure 2) enables users to customize the firm-to-firm network graph through an intuitive side panel. Here, users can select the target year, choose up to three companies for analysis, and specify the number of closely related companies to display. After configuring these settings, users can click the "SUBMIT" button to rerender the graph based on the seleceted criteria. For example, Figure 2 illustrates the top 10 relations of Alphabet Inc. and Meta Platforms. Note that the graph dynamically adjusts to display only the relations that meet the selected parameters. Users may also reset the settings to their default state by clicking the "RESET" button, which reverts to a graph showing the relations among the top 50 firms by market capitalization.

Graph Visualization Module This module (labeled as partition (2) in Figure 2) features an interactive network graph at the center of our system. Each firm is represented as a node, while the edges signify the existence of relations between firms. Key visual elements include thicker edges reflecting stronger relations and larger nodes representing firms with higher market capitalizations.

Users can hover over a node to view the name of the company it represents. The system categorizes firms by sectors using different colors, as outlined in the graph legend.[4] Selected edges are highlighted to improve visibility, allowing users to focus on specific connections. The 3D graph visualization further enhances user interactivity, enabling them to customize the view to their preferences using the control instructions displayed at the bottom center of the screen. In Figure 2, the edge between Alphabet and Meta in the graph is highlighted in red, and details of their relationship are displayed in the adjacent right panel, offering a comprehensive understanding of the selected connection.

Relation Rationale Explainer This module (labeled as partition (3) in Figure 2) provides detailed insights into the specific relation between two firms. Users can access this information by clicking on any edge in the network graph. The displayed details include: (a) the names of two connected firms, (b) the strength of their relation, (c) the ranking of the relation strength, (d) LLM-generated keywords and a summary of these MRPs, and (e) mutual risk paragraphs (MRPs) extracted from both firms'

[3]Available at https://surf-firm-risk-relations.onrender.com

[4]The categorization is based on the Global Industry Classification Standard (GICS).

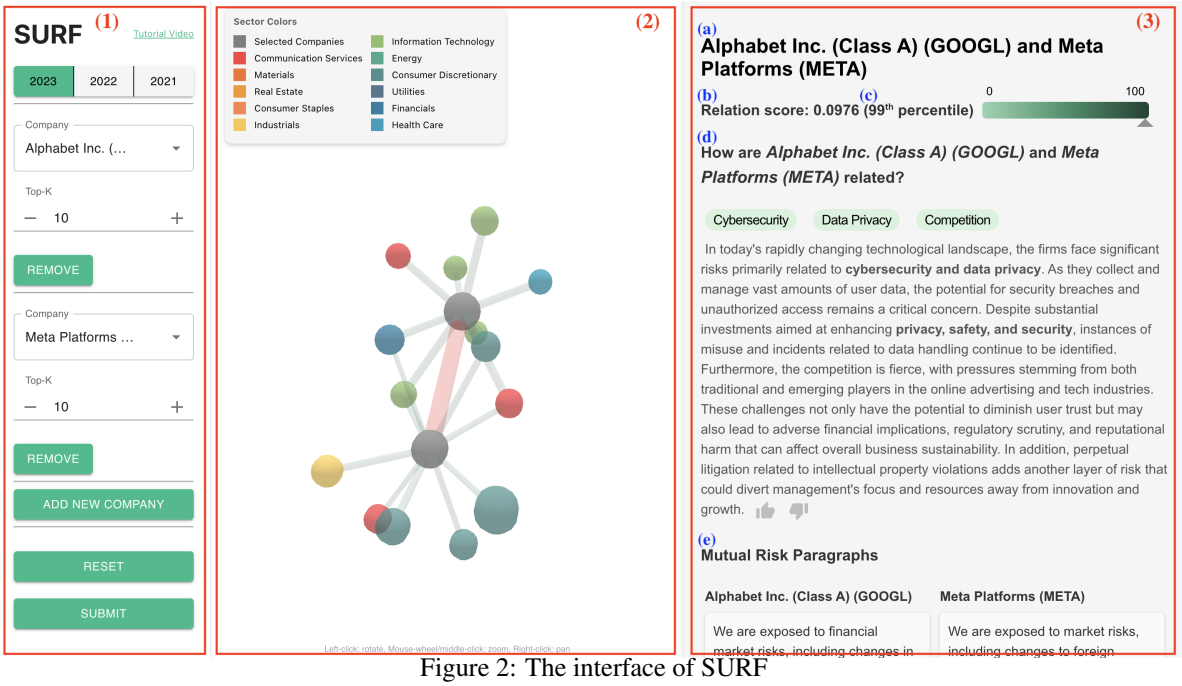

Figure 2: The interface of SURF

10-K filings.[5] Users can explore these MRPs further by clicking on a paragraph to view similar risk discussions from the other firm. To enhance user experience and minimize response time, a predefined prompt is employed (see Appendix A) to guide the large language model (LLM) in generating concise summaries of the rationale behind the relations. These summaries highlight shared risks and explain the connection between the selected firms for the given year. Figure 2 illustrates a summary that outlines shared risks and their relevance to the selected firms. We use ChatGPT,[6] based on the GPT-4o-mini architecture,[7] to generate summaries due to its high performance and cost efficiency. The initial prompt instructs the model to summarize the risks faced by the two firms using their MRPs, highlight key sections in the summary using `<strong>` HTML tags, and generate three keywords from the summary to clarify the shared risks. The initial prompt is not static. Inspired by (Sun et al., 2024), we integrate a feedback mechanism to continuously improve the quality of the generated rationales and enhance user experience. Specifically, positive and negative feedback on each rationale is collected and analyzed by the LLM to identify reasons for inaccuracies. Based on this analysis, the prompt is refined to better align with user expectations.[8] This dynamic feedback loop allows SURF to iteratively adapt the prompt, improving both clarity and relevance over time.

Key Contributions In summary, SURF advances the identification of risk relations in the following three key aspects:

1. **Explainability:** SURF offers an explainable approach to extracting risk relations from 10-K filings. Users can click on a company's MRPs to highlight the corresponding paragraphs from the other firm in red, offering a clear rationale for the identified relations. By directly examining these highlighted sections, users gain insights into how the encoder identifies the connection. Also, the proposed relation score (see (b) in Figure 2) quantifies these relations and enhances interpretability. These detailed and transparent explanations distinguish SURF from prior approaches, making it a valuable tool for understanding risk connections.
2. **User-friendly Interaction:** SURF is designed with a focus on user-friendly interaction, offering high levels of customization. Users can tailor settings to their preferences and adjust the graph presentation seamlessly by scrolling or using mouse buttons.
3. **Efficiency:** Analyzing 10-K filings is traditionally time-consuming and requires specialized financial expertise due to their extensive content

[5] We label these components in Figure 2.
[6] https://openai.com/index/chatgpt
[7] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence
[8] Details of the prompt refinement process are available in Appendix B.

and complex structure. SURF addresses these challenges by leveraging retrieval-augmented generation (RAG) techniques. It uses MRPs as inputs and use GPT-4o-mini to generate the rationales for risk relations, significantly reducing the time required for analysis. This approach lowers the barrier for non-experts to access and understand valuable insights from financial reports, making such analyses more accessible and efficient.

3 Relation Identification

Dense Encoder Training We aim to fine-tune an encoder capable of producing higher cosine similarity scores for paragraphs discussing mutual risks. However, annotating such paragraphs requires specialized financial expertise and the knowledge of the target companies. Identifying shared risks is particularly challenging due to the ambiguous language often used in regulatory documents. Even when relevant paragraphs are identified, building and scaling up a supervised dataset remains a significant challenge.

To this end, we adopt a self-supervised approach using contrastive learning (Hadsell et al., 2006; Gao et al., 2021). Specifically, we create two distinct views to fine-tune a BERT model (Devlin et al., 2019) for retrieval and build a self-supervised dataset from SEC filings. These views are designed to capture both chronological and lexical representations of the documents, as detailed below.

1. **Chronological Similarity View**: We assume that after excluding accounting-related dates, a single firm experiences only one significant event on any given day. For example, the upper section of Table 1 presents two paragraphs discussing Nvidia's termination of its purchase of Arm. Both paragraphs feature the same date format that has been highlighted for clarity.[9] Hence, positive pairs in this view are formed from two paragraphs of the same firm that contain identical date format tokens (e.g., "July 8, 2024").[10]
2. **Lexical Similarity View**: As regulatory documents, Form 10-K filings often use lexically similar words and phrases to describe analogous events. For instance, the lower section of Table 1 shows two paragraphs discussing the same risk, with overlapping words marked.[11] To leverage this characteristic, we create a lexical similarity view that captures such similarities. Positive pairs are generated from overlapping text segments within the same paragraph. For a paragraph with n words $[w_1, w_2, \ldots, w_n]$ where $n \geq 3$, we randomly select indices i and j such that $1 < i < j < n$ and form positive pairs as $([w_1, \ldots, w_j], [w_i, \ldots, w_n])$.

For each view, we generate 8,500 positive pairs for training and 1,000 positive pairs for validation. To optimize the training process, we employ the InfoNCE loss with in-batch negatives (van den Oord et al., 2019; Karpukhin et al., 2020). Training is halted early if there is no improvement in loss after five epochs. Additionally, the output vectors are normalized to unit length to facilitate the computation of cosine similarity between paragraph vectors, thereby aiding in the identification of mutual risks.

Calculation of Risk Relation Scores We leverage the trained encoder to compute paragraph embeddings, where high cosine similarity between embeddings indicates potential chronological or lexical similarities. With a pre-defined threshold ξ, a paragraph p is considered to discuss a similar risk as another paragraph q if $\cos(q, p) > \xi$.

Instead of employing the commonly used top-k retrieval method, we adopt a threshold-based approach to ensure that all retrieved paragraphs meet a minimum similarity level, thereby reliably reflecting shared risks. The threshold ξ is incrementally adjusted by 0.05 within the range of 0.6 to 0.95, with the optimal value determined to be 0.75.

Based on the above information, we identify mutual risk paragraphs (MRPs) between two firms, A and B, as paragraphs from firm A that discuss a similar risk to at least one paragraph from firm B, and vice versa. These MRPs serve as supporting evidence for shared risk relations between the firms.

Finally, the risk relation score (abbreviated as RRS, hereafter) between two firms is defined as the ratio of the number of MRPs to the total number of paragraphs from both firms. The RRS ranges from 0 (no shared risk) to 1 (maximum shared risk), with higher scores indicating stronger risk-related connections between the firms.

[9] Excerpted from NVIDIA's 10-K filing.

[10] Date format tokens are excluded from positive pairs during training and validation to prevent overfitting.

[11] Excerpted from United Parcel Service's 10-K filings.

Type	Example
Chronological Similarity	1. On February 8, 2022, NVIDIA and SoftBank Group Corp., or SoftBank, announced the termination of the Share Purchase Agreement whereby NVIDIA would have acquired Arm Limited, or Arm, from SoftBank. 2. On February 8, 2022, we announced the termination of the Share Purchase Agreement by which we would have acquired Arm due to significant regulatory challenges preventing the completion of the transaction and expect to incur a $1.36 billion charge in the first quarter of fiscal year 2023.
Lexical Similarity	1. In October 2021, we completed the acquisition of Roadie, a technology platform focused on same-day delivery services, for $586 million. The results of Roadie are reported within supply chain solutions. The acquisition did not have a material impact on our results of operations for the year. 2. Business acquisitions in October 2021, we acquired Roadie, inc. ("Roadie"), a technology platform that provides local same-day delivery with operations throughout the United States. The roadie technology platform is purpose-built to connect merchants and consumers with contract drivers to enable efficient and scalable same-day local delivery services for items that are not compatible with the UPS network.

Table 1: Inspirations for the proposed two views in contrastive learning

4 Risk Relation Identification Evaluation

Data Sources Our data consists of stock price data and Form 10-K filings from companies listed in the S&P 500 Index between 2018 and 2023.[12] Stock price data was retrieved from Yahoo Finance, and Form 10-K filings were obtained through the official API of the U.S. SEC.[13] Prior to encoding, we processed the raw text by removing all HTML tags, eXtensible Business Reporting Language (XBRL) tags, and tables. To maintain consistency, we excluded companies that underwent mergers or lacked any risk-related disclosure during this period. After applying these filters, our dataset comprises 2,136 filings from 356 companies over the six-year span.

Baselines We employ three categories of baselines to establish risk relations between companies.

- **Direct Mentions** The simplest approach to identify risk relations is counting how many times one company is mentioned in the filings of the other company. For a given company pair (A, B), the RRS is calculated as the number of times company B is mentioned in company A's 10-K filing within a given year. The resulting RRSs are positive integers, and the risk relation score matrix is asymmetric.
- **GICS Sector and GICS Industry**[14] Companies within the same sector or industry are assigned an RRS of 1, while companies in different sectors or industries receive an RRS of 0. The RRSs for this baseline are binary.
- **Other Dense Encoders** We evaluate three encoder checkpoints—DPR (Karpukhin et al., 2020), Contriever (Izacard et al., 2022), and FinBERT (Araci, 2019)—to embed paragraphs and calculate RRS as benchmarks against our encoder. DPR and Contriever are chosen as the representative retrievers for supervised and unsupervised learning, respectively. FinBERT is also included because it has been pre-trained on financial datasets.

Evaluation Metrics We hypothesize that if two firms are exposed to similar risks, they are more likely to experience a common risk event that impacts both firms' stock prices simultaneously. Based on the hypothesis, we introduce a new metric ρ —the correlation between (1) RRSs and (2) the correlation of the absolute values of daily stock returns (CAVDSR) between firms. The metric is defined as $\rho = \text{corr}(\text{RSS}, \text{CAVDSR})$.

The absolute values are used because firms can react to similar events in opposite directions. For example, the pandemic made positive impacts on the Healthcare Industry but adversely affected the Travel Industry. A higher ρ indicates that the method is more effective at capturing the shared risks between firms.

[12] The list of constituents is based on the most recent changes to the S&P 500 Index in 2023. Only companies that were constituents for all six years are included.

[13] https://www.sec.gov/search-filings/edgar-application-programming-interfaces

[14] The Global Industry Classification Standard (GICS) categorizes companies into 11 sectors and 74 industries.

	2020	2021	2022	2023
Direct Mentions	0.0124	0.0303	0.0340	0.0312
GICS Sector	0.1637	0.2845	0.2917	0.2985
GICS Industry	0.1774	0.3297	0.2929	0.3305
Contriever	0.2042	<u>0.3914</u>	<u>0.3253</u>	0.4027
DPR	<u>0.2069</u>	0.3896	0.3138	<u>0.4068</u>
FinBERT	0.1656	0.3311	0.3228	0.3079
Ours	**0.2091**	**0.4054**	**0.3373**	**0.4191**

Table 2: Correlations between RRS and CAVDSR

Performance Analysis Table 2 demonstrates the effectiveness of the proposed method for risk relations identification, where the best performance is indicated in bold, and the second-best is underlined in the table. As tabulated in the table, our two-view encoder and retrieval approach outperforms all baselines:

- **Direct Mentions** This approach struggles to capture risk relations effectively. Simply counting mentions of one firm in another's filings inadequately reflects the complex interdependencies and shared risks.
- **GICS Sector and GICS Industry** The GICS-based approaches perform better as they inherently group firms with similar risks. However, these methods overlook more nuanced lexical and semantical similarities in the filings. Classifying one company into only one industry also neglects the complexity of the business model. Compared with SURF, the classification process of GICS-based methods lacks transparency, a common limitation also found in other manual classification approaches.
- **Other Dense Encoders** DPR and Contriever show performance levels close to our encoder. However, their training process does not sufficiently emphasize chronological similarity, limiting their effectiveness. Surprisingly, FinBERT, a finance-specific encoder, performs the worst among the four retrieval-based approaches (Ours, Contriever, DPR, and FinBERT). This result may stem from the fact that FinBERT is not explicitly trained for retrieval tasks.

5 Case Study

To demonstrate SURF's ability to uncover latent relations, we present a case study on Nvidia (NVDA), a leader in high-performance GPUs and AI processing, and Wabtec (WAB), a pioneer in advanced rail and transit solutions. In 2022, SURF identified a strong risk relation between these companies, ranking in the 95th percentile among 63,190 analyzed relations. While this connection may seem unexpected, a deeper analysis reveals that SURF detected underlying links, particularly concerning supply chain disruptions.

One key risk highlighted in the LLM-generated summary indicates that the COVID-19 pandemic exacerbated operational disruptions for both firms, notably through supply chain constraints. Similar insights appear in their MRPs, reinforcing this shared risk which is further supported by news reports in 2022:

- Reuters reported that Nvidia was impacted by the global chip shortage which kept GPU prices high. As supply constraints eased and prices fell, analysts viewed it as a negative market signal that would lead to a decline in Nvidia's stock.[15]
- In Q1 2022, an equity research institution noted that "supply-chain disruptions (including higher commodity costs and shortages of components, chips, and labor) might have also dampened Wabtec's first-quarter performance."[16]

These events underscore the critical role of chip supply chain disruptions as a shared risk factor for both companies and validate SURF's effectiveness in identifying hidden interdependencies.

6 Conclusions

In this paper, we introduce SURF, a system designed to visualize, quantify, and interpret the risk relations between companies. SURF provides users with an intuitive interface to understand how companies are connected through shared risks, why these risks exist, and how they may impact both companies. Both experts and non-experts can leverage SURF to make more informed decisions.

Beyond the system design, our technical contribution lies in the development of an encoder that effectively extracts latent information from the text of financial reports. In our experiments, the encoder consistently outperforms all baseline methods. This technical advancement highlights SURF's ability to unravel the complexities of financial documents, making them more transparent and accessible to a broad spectrum of users.

Future Work We leverage 10-K filings as a robust and reliable data source. These filings are

[15] https://www.reuters.com/technology/graphic-chip-price-drop-raises-questions-whether-end-shortage-is-sight-2022-04-25/

[16] https://www.nasdaq.com/articles/whats-in-the-offing-for-wabtec-wab-this-earnings-season

audited and provide comprehensive narratives of events with material impacts, offering a strong foundation for our analysis. To further enhance the scope of our work, we plan to integrate additional data sources, such as news articles and analyst reports, to complement the 10-K filings and provide users with more timely and diverse insights.

In addition, our method and system open new avenues for future research. For example, the practice of disclosing important dates in reports is not unique to the finance field. The chronological similarity view we developed has the potential to discover connections/relations between entities (e.g., geopolitical risks) across various types of corpora. Similarly, the rationale explainer in SURF can be adapted to other systems to enhance transparency in complex or opaque content. We aim to build on our current system and methodology to expand their application to different fields, fostering broader interdisciplinary insights into future research.

7 Ethics Statement

The content generated by SURF is provided solely for reference purposes and is not intended to serve as the basis for any investment decisions. It does not reflect the views of the authors or their affiliated institutions. Users are advised that LLM-generated content is not a substitute for professional advice.

References

Dogu Araci. 2019. FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv:1908.10063*.

Philippe Bracke, Anupam Datta, Carsten Jung, and Shayak Sen. 2019. Machine learning explainability in finance: An application to default risk analysis. *Bank of England Working Paper No. 816*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1735–1742.

Andreas GF Hoepner, David McMillan, Andrew Vivian, and Chardin Wese Simen. 2021. Significance, relevance and explainability in the machine learning age: An econometrics and financial data science perspective. *The European Journal of Finance*, 27(1-2):1–7.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wentau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6769–6781.

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

Zhu Sun, Hongyang Liu, Xinghua Qu, Kaidong Feng, Yan Wang, and Yew Soon Ong. 2024. Large language models for intent-driven session recommendations. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 324–334.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding. *arXiv:1807.03748*.

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

A Prompts for Summary Generation

Prompt 1: Initial Task Description

The following two lists represent similar risks from two different firms. Please summarize the key risks into a coherent and concise paragraph, combining both lists. Within the summary, use the HTML `<strong>` tag to highlight up to three important phrases or sentences. Additionally, select up to three relevant keywords from the combined text that represent central concepts or themes. The output should start with "Keywords:" followed by the selected keywords, separated by commas.

After the keywords, leave a blank line before providing the summary. Ensure the format follows this structure: "Keywords: RiskA, RiskB, RiskC" followed by the paragraph with important phrases or sentences highlighted in bold using the tag.

{MRPs_from_the_1st_firm}

{MRPs_from_the_2nd_firm}

B Prompts for prompt refinement with feedback

Prompt 1: Inferring Reasons for Errors

I'm trying to write a zero-shot paragraph summarization prompt.
My current prompt is {prompt}.
But this prompt gets the following example wrong: {error_case}, give {N_r} reasons why the prompt could have gotten this example wrong.
Wrap each reason with <START> and <END>.

Prompt 2: Refining Prompts with Reasons

I'm trying to write a zero-shot paragraph summarization prompt.
My current prompt is {prompt}.
But this prompt gets the following example wrong: {error_case}.
Based on the example, the problem with this prompt is that {reasons}.
Based on the above information, please write one improved prompt. The prompt is wrapped with <START> and <END>.
The new prompt is: