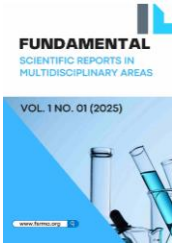
Fundamental Scientific Reports in Multidisciplinary Areas

Vol: 02 Issue:03 2026

E-ISSN: 3134-7517 P-ISSN: 3134-7509



ARTICLE HISTORY

Received: [10 October 2026]
Revised: [17 October 2026]
Accepted: [24 October 2026]

Published online: [31 October 2026]

KEYWORDS

Large Language Model; Retrieval-Augmented Generation; Dynamic Risk Diffusion; Private Enterprise Distress Prediction; Risk-Aware Decision Support; Financial Artificial Intelligence

LICENSE

This article is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0). You may share and adapt the material in any medium or format, with appropriate credit to the original author(s) and a link to the licence.
https://creativecommons.org/licenses/by/4.0/

COPYRIGHT

© 2026 The Author(s). Authors retain copyright; FSRMA holds a non-exclusive licence to publish.

ORIGINAL RESEARCH ARTICLE

R2D-LLM: Retrieval-Augmented Large Language Model Driven Dynamic Risk Diffusion Network for Private Enterprise Distress Prediction and Risk-Aware Investment Decision Support

Wenxuan Liu [1], Chongli Wu [2], Yuchen Gan [3]

[1] *Computer Science, University of California, Los Angeles, Los Angeles, CA, USA*
[2] *Master in Applied Finance, Pepperdine University, Malibu, CA, USA*
[3] *Master of Science, Columbia University, New York, NY, USA*
[1] *vincentliux@g.ucla.edu*
[2] *Chongli.wu@yahoo.com*
[3] *yg2680@columbia.edu*

ABSTRACT

Private enterprises in venture capital (VC) and private credit markets are difficult to assess because limited disclosure, incomplete credit histories, and rapidly changing operating conditions restrict the usefulness of conventional ratio-based models. Important warning signals are often dispersed across financing announcements, regulatory disclosures, business reports, and market news, while enterprise distress typically develops through interacting events rather than a single observation. This study proposes R2D-LLM, a retrieval-augmented large language model driven dynamic risk diffusion framework for private enterprise distress prediction and risk-aware investment decision support. The framework uses a RAG-enhanced LLM to convert heterogeneous documents into structured risk events containing category, severity, timestamp, semantic description, and confidence information. A dynamic risk event graph and Risk Diffusion Transformer then model temporal dependencies, risk accumulation, cross-category propagation, and attenuation. Lifecycle embeddings and enterprise-level financial and contextual features are integrated to estimate twelve-month distress risk across different financing stages. On the reported multi-source dataset containing 12,846 enterprises, 186,520 documents, and 94,317 extracted risk events, R2D-LLM achieves 92.8% accuracy, a 90.6% F1-score, and an AUC of 0.936, outperforming the evaluated machine-learning, deep-learning, Transformer, and financial-LLM baselines. Reported ablation results suggest that retrieval-enhanced event extraction and dynamic diffusion modeling provide the largest contributions. These findings support R2D-LLM as a dynamic enterprise risk-assessment framework and provide a basis for subsequent validation of downstream investment strategies.

1. Introduction

Private enterprises, particularly startups and growth-stage companies, have become important drivers of technological innovation and economic development. Venture capital (VC) and private credit markets provide essential financing channels for these enterprises; however, investment decision-making in private markets remains highly challenging due to severe information asymmetry, limited financial disclosure, and rapidly changing business environments. Unlike publicly listed companies, private enterprises often lack comprehensive financial reporting and stable credit histories, making traditional credit risk assessment methods based on financial ratios, leverage indicators, and historical default records insufficient for early distress detection. The twelve-month distress outcome is defined in Section 3.5. In practice, critical risk signals frequently emerge from heterogeneous unstructured information, including financing announcements, regulatory documents, founder communications, business reports, customer feedback, and market news. These textual sources contain valuable but implicit indicators related to liquidity pressure, financing difficulties, technological uncertainty, management instability, and regulatory exposure.

Recent advances in large language models (LLMs) have demonstrated strong capabilities in semantic understanding, reasoning, and information extraction, providing new opportunities for intelligent financial risk analysis. However, existing LLM-based financial risk models mainly focus on static text classification or sentiment analysis, while overlooking the dynamic evolution mechanisms of enterprise risks. In private markets, enterprise distress is rarely caused by an isolated event; instead, it typically results from a continuous process of risk accumulation and propagation. For example, technological setbacks may trigger market confidence decline, leading to customer loss, financing difficulties, and ultimately liquidity crises. Therefore, accurately predicting enterprise distress requires not only extracting risk-related information from textual data but also modeling the temporal dependencies, self-excitation effects, and cross-category diffusion patterns among risk events. Based on these observations, the central question is whether unstructured textual signals, when represented as time-ordered and relational risk events, provide predictive information beyond static financial and textual features. More specifically, the study examines whether retrieval-enhanced event extraction, dynamic risk diffusion modeling, and lifecycle information can jointly improve twelve-month distress prediction for private enterprises with limited disclosure. To address these challenges, this study proposes R2D-LLM (Retrieval-Augmented Large Language Model Driven Dynamic Risk Diffusion Network), an intelligent framework for private enterprise distress prediction and downstream risk-aware decision support. R2D-LLM integrates retrieval-augmented large language models with dynamic risk diffusion modeling to establish an end-to-end risk intelligence framework. Specifically, the framework first employs a retrieval-enhanced LLM module to extract structured risk events from multi-source enterprise documents by incorporating external risk knowledge bases. Then, a dynamic risk event graph is constructed to represent relationships among enterprises, risk factors, and temporal events. Furthermore, a Risk Diffusion Transformer is developed by combining temporal attention mechanisms with neural point process modeling to capture risk accumulation, propagation, and attenuation processes. Finally, a lifecycle-aware distress prediction module is developed and a downstream decision-support extension is outlined to support VC and private credit decision-making across different enterprise development stages.

The study makes three contributions:

(1) A retrieval-enhanced risk event representation approach is developed to transform heterogeneous private-enterprise documents into structured events containing risk category, severity, temporal information, semantic description, and confidence estimates. This design enables weak and implicit textual warning signals to be incorporated into enterprise distress analysis.

(2) A dynamic enterprise risk modeling architecture is proposed by combining a Dynamic Risk Event Graph, temporal attention, neural point-process modeling, and lifecycle information. The framework represents distress as a process of risk accumulation and cross-category propagation rather than as a static classification outcome, while accounting for heterogeneous risk patterns across financing stages.

© 2026 The Author(s). Licensed under CC BY 4.0

(3) The framework is evaluated on a multi-source private-enterprise dataset containing 12,846 enterprises, 186,520 documents, and 94,317 extracted risk events. Comparative and ablation results indicate the predictive value of retrieval-enhanced event extraction, dynamic diffusion modeling, and lifecycle-aware representation. A downstream risk-aware decision-support layer is further outlined to connect distress estimates with candidate investment, holding, and exit actions.

The reported experiments evaluate distress prediction. Investment performance is outside the scope of the current evaluation and requires separate validation.

2.Literature Review

Recent advances in artificial intelligence (AI), natural language processing (NLP), and financial machine learning have significantly transformed enterprise risk assessment and investment decision-making. Traditional credit risk models mainly rely on structured financial variables, while emerging approaches attempt to exploit unstructured textual information, temporal risk evolution patterns, and large-scale financial knowledge. In particular, the rapid development of large language models (LLMs) has provided new opportunities for extracting implicit risk signals from enterprise documents, whereas temporal event modeling techniques have enabled dynamic analysis of risk propagation processes. This section reviews four research directions closely related to this study, including LLM-driven financial intelligence, retrieval-augmented financial analysis, dynamic risk event modeling, and AI-based credit risk prediction and investment optimization.

2.1 Large Language Models for Financial Risk Analysis

Large language models have recently demonstrated remarkable capabilities in financial information understanding, reasoning, and knowledge extraction. Unlike traditional natural language processing models that mainly focus on sentiment classification or entity recognition, LLMs can interpret complex financial narratives and identify implicit relationships among companies, markets, and risk factors.

Brown et al. [1] introduced GPT-3, demonstrating that large-scale pretrained language models can achieve strong few-shot learning performance across diverse NLP tasks. This work established the foundation for applying LLMs to domain-specific knowledge understanding. However, general-purpose LLMs lack specialized financial knowledge and may generate inaccurate interpretations in high-risk financial scenarios. To address this limitation, Yang et al. [2] proposed FinGPT, an open-source financial large language model designed for financial applications. FinGPT introduced a data-centric training paradigm with financial instruction tuning and demonstrated promising performance in financial analysis tasks, including market interpretation and financial information processing. Similarly, Wu et al. [3] developed BloombergGPT, a large language model specifically trained on extensive financial corpora, showing that domain-specific pretraining can significantly improve financial language understanding.

Although these studies demonstrate the potential of LLMs in finance, existing approaches mainly focus on financial question answering, sentiment analysis, and market information extraction. They rarely consider the dynamic evolution of enterprise risks or the propagation mechanisms among different risk events. Therefore, directly applying existing financial LLMs to private enterprise distress prediction remains challenging because startup risks are usually hidden in scattered textual signals rather than explicit financial indicators.

2.2 Retrieval-Augmented Generation for Financial Knowledge Enhancement

Retrieval-Augmented Generation (RAG) has recently emerged as an effective approach for improving the reliability and domain adaptability of LLM-based systems. By integrating external knowledge retrieval with language generation, RAG enables models to access updated and domain-specific information while reducing hallucination problems.

Lewis et al. [4] proposed the original Retrieval-Augmented Generation framework, which combines neural retrieval models with pretrained language models to incorporate external knowledge during generation. Their work demonstrated that retrieval mechanisms can substantially improve factual accuracy and knowledge-intensive reasoning tasks.

© 2026 The Author(s). Licensed under CC BY 4.0

In financial applications, RAG offers a way to supplement model-generated analysis with retrieved evidence. Gao et al. [5] provide a survey of retrieval-augmented generation for large language models.
RAG supplies external context, but retrieval alone does not represent the order or interaction of enterprise risk events. Knowledge-graph-based semantic retrieval has also been explored in other domains [6]; such work provides architectural context rather than evidence of improved financial-risk extraction. The present study combines retrieved context with an explicit temporal event representation.

2.3 Dynamic Event Propagation and Temporal Risk Modeling

Enterprise distress can be studied as a sequence of interacting events. Temporal point processes and neural event models provide tools for representing event timing and cross-category dependence rather than treating each observation in isolation.
Hawkes Process, originally introduced by Hawkes [7], provides a mathematical framework for modeling self-exciting event sequences, where previous events can increase the probability of future events. This mechanism is particularly suitable for financial risk propagation because one risk event may trigger subsequent adverse events. Mei and Eisner [8] proposed the Neural Hawkes Process, which integrates recurrent neural networks with temporal point processes to capture complex event dependencies. Their model enables flexible modeling of event intensity evolution and demonstrates improved predictive capability compared with traditional Hawkes models. Furthermore, Zuo et al. [9] introduced the Transformer Hawkes Process, which incorporates self-attention mechanisms into Hawkes modeling to capture both short-term and long-term event dependencies. The model achieved superior performance in various event prediction tasks by effectively learning temporal relationships. Recent studies have further explored graph-based temporal point processes and uncertainty-aware event prediction. For example, Yoon et al. [10] proposed a graph recurrent temporal point process framework to model multivariate event interactions, demonstrating the effectiveness of graph structures for capturing dependencies among heterogeneous events. Although these methods provide powerful tools for temporal event modeling, most existing approaches focus on social networks, user behavior, or general event prediction. Their applications to private enterprise distress prediction remain limited. In particular, existing temporal models lack semantic understanding of risk events extracted from financial documents, motivating the integration of LLM-based risk perception with dynamic event diffusion modeling.

2.4 AI-based Credit Risk Prediction and Investment Decision Optimization

Artificial intelligence-based credit risk assessment has been extensively studied in financial risk management. Traditional machine learning methods, including logistic regression, random forests, and gradient boosting models, have been widely applied for default prediction using structured financial indicators. However, these approaches usually require sufficient historical financial records and cannot effectively handle early-stage enterprises with limited disclosure. Recent studies have explored deep learning approaches for credit risk prediction. Mestiri et al. [11] proposed a deep neural network-based credit scoring model that improved nonlinear feature representation compared with traditional statistical methods. Mojdehi et al. [12] further introduced graph neural network-based credit risk models to capture relational information among financial entities. However, most existing AI-based credit risk models focus on publicly available financial institutions or consumer credit scenarios. Private enterprises present unique challenges because their risks are strongly influenced by qualitative factors, including management capability, technological uncertainty, investor confidence, and regulatory changes. More recently, LLM-based financial agents and intelligent decision systems have attracted increasing attention. Financial foundation models such as FinGPT demonstrate the possibility of using LLMs for automated financial analysis and decision support [2]. Nevertheless, existing investment intelligence systems mainly provide information summarization or recommendation generation rather than explicitly optimizing decisions based on dynamic risk evolution. The present study therefore examines distress prediction from retrieved evidence and dynamic risk events, while treating investment decision support as a downstream application rather than an empirically validated contribution.

© 2026 The Author(s). Licensed under CC BY 4.0

3.Methodology

3.1Overview of the Proposed R2D-LLM Framework

This study proposes R2D-LLM (Retrieval-Augmented Large Language Model Driven Dynamic Risk Diffusion Network), a unified framework that integrates large language model-based risk perception, dynamic event propagation modeling, enterprise distress prediction, and downstream decision support. The fundamental motivation of R2D-LLM is that private enterprise distress is not an isolated financial outcome but an evolving process driven by the accumulation and interaction of heterogeneous risk events. Unlike traditional credit risk models that mainly rely on structured financial indicators, R2D-LLM utilizes large-scale textual information generated throughout enterprise development cycles, including financing announcements, regulatory reports, business descriptions, investor communications, and market news, to identify early-stage risk signals. The proposed framework consists of four interconnected components: (1) a Retrieval-Augmented Generation (RAG) enhanced large language model risk perception module for extracting structured risk events from unstructured enterprise documents; (2) a Dynamic Risk Event Graph construction module for representing temporal and relational dependencies among enterprise risks; (3) a Risk Diffusion Transformer module for learning dynamic risk propagation patterns based on temporal attention and neural point processes; and (4) a lifecycle-aware distress prediction module with an outlined decision-support extension for generating risk scores and supporting VC/private credit decision-making. The overall objective of R2D-LLM is to transform heterogeneous enterprise information into dynamic risk intelligence and provide risk estimates for subsequent investment review.

Given an enterprise document collection:

$$D = \{d_1, d_2, \ldots, d_N\}$$

the framework first extracts a sequence of risk events:

$$E = \{e_1, e_2, \ldots, e_T\}$$

where each event contains semantic, temporal, and risk-related attributes. The final prediction task can be formulated as estimating the probability of future enterprise distress:

$$P(Y_{t+\Delta} = 1 \mid E_{1:t}, X_t)$$

where Y represents enterprise distress status, $E_{1:t}$ denotes historical risk events, and X_t represents enterprise contextual information. Figure 1 links retrieved enterprise evidence to structured risk events, a dynamic graph, temporal diffusion modeling, and lifecycle-aware prediction. The event layer preserves category, severity, time, description, and confidence information before graph and temporal aggregation. Section 4 evaluates twelve-month distress prediction and outcome-severity estimation. The subsequent decision layer is an outlined extension, not a validated investment strategy.

© 2026 The Author(s). Licensed under CC BY 4.0

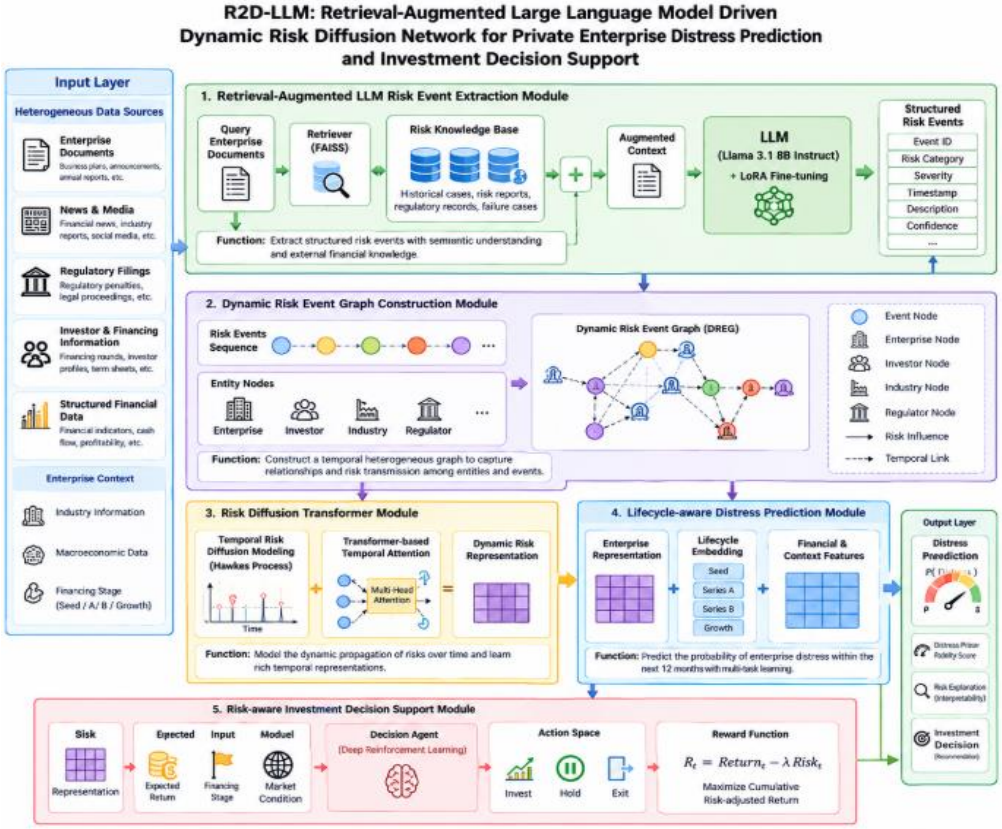


Figure 1R2D-LLM architecture. The prediction path is evaluated in Section 4; the downstream risk-aware decision-support layer is a conceptual extension.

Based on the overall architecture shown in Figure 1, the following subsections explain how unstructured enterprise information is transformed into structured risk events, organized into a dynamic graph, and used for temporal distress prediction.

3.2 Retrieval-Augmented Large Language Model-Based Risk Event Extraction

The extraction stage converts enterprise documents into structured, temporally grounded risk events. RAG supplies relevant domain context before the LLM produces an event record.

A domain-specific risk knowledge base is first constructed from historical enterprise distress cases, regulatory enforcement records, financing failure reports, credit-risk cases, industry reports, and venture capital analysis materials. Each knowledge segment x_j is transformed into a dense representation through an embedding function:

$$z_j = f_\theta(x_j)$$

where $f_\theta(\cdot)$ denotes the embedding encoder. For an input enterprise document q, its embedding z_q is compared with the knowledge representations using cosine similarity:

$$\mathrm{sim}(q, x_j) = \frac{z_q^T z_j}{\|z_q\|\|z_j\|}$$

The retrieval module then selects the top-K most relevant knowledge entries:

$$K_q = \{k_1, k_2, \ldots, k_\mathrm{K}\}$$

where K denotes the number of retrieved knowledge segments and k_i represents the i-th retrieved segment. This notation explicitly distinguishes the retrieval size K from an individual retrieved item k_i. The retrieved evidence is concatenated with the original enterprise document and supplied to the LLM as contextual

© 2026 The Author(s). Licensed under CC BY 4.0

evidence. This design allows the model to ground its risk interpretation in domain-specific historical knowledge rather than relying exclusively on its parametric knowledge.

The extraction prompt requests the enterprise, risk category, description, occurrence time, severity, and confidence. It distinguishes documented facts from inferred risks and normalizes dates and severity according to the coding scheme. The structured event is represented as:

$$e_i = (c_i, s_i, t_i, r_i, v_i)$$

where c_i denotes the risk category, s_i represents normalized risk severity, t_i denotes the event timestamp, r_i contains the semantic description of the event, and $v_i \in [0,1]$ represents the model-estimated confidence of the extracted event. The risk taxonomy includes liquidity risk, financing risk, technological risk, market risk, management risk, and regulatory risk. Severity is normalized to the interval $[0,1]$, with larger values indicating greater potential impact on enterprise stability.

Importantly, the confidence score v_i is not treated as an unused auxiliary attribute. It is incorporated into subsequent dynamic risk diffusion modeling as an event reliability weight. Specifically, the effective event weight is defined as:

$$\omega_i = v_i \cdot s_i$$

where ω_i combines assigned severity and model-estimated confidence for event i. Low-confidence events receive lower weights; this weighting does not establish that confidence is calibrated to extraction accuracy. The weighted event representation is:

$$\tilde{h}_i = \omega_i \cdot h_i$$

where h_i is the semantic representation generated from the LLM. The confidence-weighted representation $\tilde{h}_i$ is used as the event-level representation when constructing the Dynamic Risk Event Graph (DREG), and the resulting graph representation is subsequently provided to the Risk Diffusion Transformer (RDT) together with the temporal event representations. Thus, the confidence weighting affects the downstream dynamic-risk representation rather than forming an independent prediction branch. These confidence-weighted event representations are subsequently passed to the Dynamic Risk Event Graph and Risk Diffusion Transformer described in Sections 3.3 and 3.4. This design passes model-estimated confidence to the downstream representation; its reliability requires independent extraction validation.

For example, when an enterprise document states that "the company postponed its Series B financing because of declining customer acquisition efficiency," the LLM identifies financing difficulty and market deterioration as related risk events, assigns the corresponding occurrence time and severity, and estimates an extraction confidence v_i. Rather than treating the sentence as a static risk label, R2D-LLM converts it into a temporally grounded event that can subsequently influence the intensity of financing, market, and liquidity risks. This enables the framework to represent risk accumulation and cross-category propagation more faithfully.

3.3 Dynamic Risk Event Graph Construction

The Dynamic Risk Event Graph (DREG) represents temporal and relational dependencies among enterprises, risk factors, and historical events. It allows technological, market, financing, and liquidity events to be modeled jointly rather than as independent observations.

The dynamic graph at time t is defined as:

$$G_t = (V_t, A_t)$$

where V_t is the node set and A_t represents temporal relationships. Nodes comprise enterprises, risk events, investors, and industry factors. Risk-event attributes follow the representation defined in Section 3.2.

The adjacency matrix A_t describes risk influence relationships among events:

© 2026 The Author(s). Licensed under CC BY 4.0

$$A_{ij} = w_{ij}$$

where w_{ij} parameterizes the modeled association strength of risk event i on event j. Here, risk diffusion denotes learned temporal and relational dependence, not identified causal transmission. Graph edges and attention weights are not estimates of intervention effects.

3.4 Risk Diffusion Transformer for Dynamic Risk Evolution Modeling

The core innovation of R2D-LLM is the proposed Risk Diffusion Transformer (RDT), which combines transformer-based temporal representation learning with neural temporal point process modeling to capture the evolution mechanism of enterprise risks.

Traditional risk prediction models usually assume that observations are independent, while real-world enterprise risks exhibit self-excitation and cascading effects. Therefore, RDT models risk occurrence as a dynamic event generation process. For each risk type i, the occurrence intensity is defined as:

$$\lambda_i(t) = \mu_i + \sum_{j=1}^{N} \omega_{ij} g(t - t_j)$$

Where μ_i denotes the baseline risk intensity, ω_{ij} represents the propagation influence from risk event j to event i, and $g(t - t_j)$ is a temporal decay function.

To enhance the capability of capturing long-term dependencies, transformer self-attention is introduced into the risk diffusion process:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}})V$$

where $Q, K,$ and V are query, key, and value matrices generated from risk event embeddings. Through temporal attention, the model can identify critical historical events that contribute most significantly to future enterprise distress.

The final hidden representation is obtained by integrating graph risk features and temporal diffusion features:

$$H_t = \text{Transformer}(E_t) + \text{Graph}(G_t) + Attention(E_t)$$

which provides a comprehensive representation of enterprise risk evolution trajectories. The three components are aligned at the event-time representation level and transformed into compatible hidden dimensions by their respective representation functions before fusion. The graph branch provides relational risk features, the diffusion branch provides temporal intensity-related features, and the attention branch provides long-range temporal dependency features; these three representations are then fused through the additive operation shown above to form the unified dynamic risk representation H_t.

The intensity component does not introduce a separate training objective. It's optimized jointly through the existing multi-task prediction losses defined in Section 3.5, namely $\mathcal{L}_{distress}$, $\mathcal{L}_{default}$, and $\mathcal{L}_{risk}$.

3.5 Lifecycle-aware Enterprise Distress Prediction

Enterprise risk characteristics vary substantially across different financing stages. Early-stage enterprises are generally more exposed to technological uncertainty, market validation difficulties, and financing constraints, whereas growth-stage enterprises are more likely to face liquidity pressure, operational complexity, debt repayment burdens, and regulatory risks. Therefore, directly applying a unified distress prediction model to enterprises at different lifecycle stages may overlook stage-specific risk characteristics. To address this issue, R2D-LLM introduces a lifecycle-aware prediction module that explicitly incorporates financing-stage information into enterprise risk representation and multi-task prediction.

The financing stage is represented through a learnable embedding:

© 2026 The Author(s). Licensed under CC BY 4.0

$$S_f = \text{Embedding}(\text{stage})$$

where $\text{stage} \in \{\text{Seed}, \text{Series A}, \text{Series B}, \text{Growth}\}$. The final enterprise representation is obtained by integrating dynamic risk evolution features, lifecycle information, and structured enterprise-level characteristics:

$$Z = [H_t, S_f, X_f]$$

where H_t denotes the dynamic risk representation generated by the risk diffusion network, S_f represents the financing-stage embedding, and X_f contains additional enterprise-level financial and market features.

To distinguish different manifestations of enterprise deterioration, R2D-LLM adopts a multi-task prediction strategy to simultaneously estimate distress probability, default probability, and overall risk level. The primary distress prediction is formulated as:

$$\hat{y}_d = \sigma(W_d Z + b_d)$$

where $\sigma(\cdot)$ denotes the sigmoid activation function, and $\hat{y}_d$ represents the predicted probability of enterprise distress.

Distress denotes financing failure, a severe liquidity crisis, default, or significant operational decline within the subsequent twelve months. Default refers specifically to failure to meet contractual debt or repayment obligations and is therefore included in the distress label. This is a relationship between outcome labels; the separate sigmoid heads do not by themselves impose an ordering on the predicted probabilities.

The default prediction task is formulated as:

$$\hat{y}_{\text{def}} = \sigma(W_{\text{def}} Z + b_{\text{def}})$$

while the risk-level prediction is defined as:

$$\hat{y}_{\text{risk}} = \text{Softmax}(W_{\text{risk}} Z + b_{\text{risk}})$$

where $\hat{y}_{\text{def}}$ denotes the predicted default probability and $\hat{y}_{\text{risk}}$represents the probability distribution over predefined enterprise risk levels.

The overall multi-task optimization objective is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{distress}} + \alpha \mathcal{L}_{\text{default}} + \beta \mathcal{L}_{\text{risk}}$$

where $\mathcal{L}_{\text{distress}}$, $\mathcal{L}_{\text{default}}$, and $\mathcal{L}_{\text{risk}}$ denote the binary cross-entropy loss for distress prediction, the binary cross-entropy loss for default prediction, and the cross-entropy loss for risk-level classification, respectively. The distress prediction task is assigned the largest contribution because it constitutes the primary objective of the proposed framework. The default and risk-level tasks serve as auxiliary objectives that provide additional supervision regarding credit failure and risk severity.

In the final implementation, the loss weights are set to $\alpha = 0.5, \beta = 0.3$. These values were selected based on validation-set performance through a small-scale hyperparameter search over candidate weight combinations. The selected configuration provides a suitable balance between the primary distress prediction task and the two auxiliary tasks, while avoiding excessive optimization toward the relatively sparse default events. During model training, the three task losses are jointly optimized, allowing shared enterprise representations to learn both general distress patterns and more specific credit-risk signals.

Distress classification is the primary task; default and risk-level prediction supply auxiliary supervision. Tables 2–3 report distress metrics and expected-severity error, but do not separately establish the accuracy of the default head. These outputs can inform the downstream decision-support extension.

3.6 Risk-Aware Decision Support (Conceptual Extension)

A downstream decision-support extension may combine predicted risk states, expected-return inputs, and financing-stage information to inform investment, holding, or exit decisions. Financial guidance systems [13] provide related context. Here this extension is conceptual: Section 4 evaluates distress prediction and

© 2026 The Author(s). Licensed under CC BY 4.0

outcome severity, not a trained investment policy or realized returns. Accordingly, the manuscript makes no empirical claim of investment-policy improvement.

4. Experiment

4.1 Dataset Preparation

To evaluate the effectiveness of R2D-LLM (Retrieval-Augmented Large Language Model Driven Dynamic Risk Diffusion Network) for private enterprise distress prediction, this study constructs a multi-source private enterprise risk dataset by integrating structured financial information with heterogeneous unstructured enterprise-related documents. The dataset is collected from publicly available venture capital databases, company announcements, regulatory disclosure platforms, financial news repositories, and industry research reports. The data collection period covers 2018–2025, including startup companies from different industries and financing stages, such as Seed, Series A, Series B, and Growth stages. The final dataset contains 12,846 private enterprises, 186,520 enterprise-related documents, and 94,317 risk events extracted through the proposed RAG-enhanced LLM risk extraction module.

Table 1 shows that the dataset combines static enterprise characteristics with dynamic event-level information. Industry type and company age describe background heterogeneity, while funding stage and funding amount indicate lifecycle position and dependence on external capital. Revenue growth and cash-flow status provide a structured view of operating and liquidity conditions. In contrast, risk category, severity, and timestamp represent the content, intensity, and sequence of adverse events extracted from enterprise documents. This design is consistent with the central assumption of R2D-LLM: two firms with similar current financial profiles may have different future distress risks because the types, severity, and timing of their risk events differ. Funding-stage information also provides the basis for lifecycle-aware prediction across Seed, Series A, Series B, and Growth stages.

Table 1Enterprise-, financing-, financial-, event-, and temporal-level features in the study dataset.

Feature Category	Feature	Description
Enterprise Information	Industry Type	Business sector of the enterprise
Enterprise Information	Company Age	Years since establishment
Financing Information	Funding Stage	Seed, Series A, Series B, or Growth stage
Financing Information	Funding Amount	Total capital raised during financing rounds
Financial Indicators	Revenue Growth	Annual revenue growth rate
Financial Indicators	Cash Flow Status	Enterprise liquidity condition
Risk Events	Risk Category	Liquidity, financing, technology, market, management, or regulatory risk
Risk Events	Risk Severity	Impact level of identified risk events
Temporal Features	Event Timestamp	Occurrence time of each risk event
Target Variable	Distress Label	Whether the enterprise experiences financial distress

The target is enterprise distress within the subsequent twelve months, using the outcome categories defined in Section 3.5. The reported dataset split is chronological, with 70% for training, 15% for validation, and 15% for testing. This evaluation concerns distress forecasting rather than investment strategy performance.

© 2026 The Author(s). Licensed under CC BY 4.0

4.2 Experimental Setup

The experiments use the chronological split described in Section 4.1. The reported environment is PyTorch 2.1 on an NVIDIA RTX 4090 GPU (24 GB), an Intel Xeon Gold 6338 CPU, and 128 GB RAM. Retrieval uses BGE-large embeddings and FAISS, and the LLM produces the event records described in Section 3.2. The reported AdamW learning rate is 2×10^{-5}, with batch size 32 and 100 training epochs.

For risk-event extraction, the framework uses the Llama 3.1 8B Instruct checkpoint as the base large language model. The LLM is adapted using parameter-efficient instruction tuning with Low-Rank Adaptation (LoRA) rather than full-parameter fine-tuning. This configuration reduces the number of trainable parameters while preserving the pretrained language knowledge of the base model. For retrieval-augmented event extraction, the retriever selects the $top-(K) = 5$ most relevant knowledge segments for each input document according to cosine similarity between the document embedding and knowledge-base embeddings.

4.3 Evaluation Metrics

The proposed R2D-LLM framework is evaluated from multiple perspectives, including binary distress classification, risk discrimination, and continuous risk estimation. Since enterprise distress prediction is primarily formulated as a binary classification task, Accuracy, Precision, Recall, and F1-score are adopted to evaluate classification performance. Accuracy measures the overall proportion of correctly classified enterprises, while Precision and Recall evaluate the capability of identifying distressed enterprises and reducing missed distress cases, respectively. The F1-score is emphasized as the primary classification metric because private enterprise risk datasets generally exhibit class imbalance, making it more informative than Accuracy alone. In addition, the Area Under the Receiver Operating Characteristic Curve (AUC) is used to evaluate the model's discrimination capability across different classification thresholds.

The additional evaluation target is a normalized outcome-severity score, not a ground-truth distress probability. It encodes the most severe observed outcome during the same twelve-month window. The risk-level head produces a continuous expected-severity estimate for comparison with this ordinal target.

Specifically, each enterprise is assigned a normalized outcome-severity score $R_i \in [0,1]$ according to the most severe observed outcome within the 12-month horizon. Five outcome levels are defined: non-distressed enterprises are assigned $R_i = 0$; enterprises experiencing significant operational deterioration are assigned $R_i = 0.25$; enterprises experiencing financing failure are assigned $R_i = 0.50$; enterprises experiencing severe liquidity crises are assigned $R_i = 0.75$; and enterprises experiencing a default event are assigned $R_i = 1.00$. When multiple adverse outcomes occur for the same enterprise during the observation period, the maximum applicable severity level is used. Therefore, default represents the most severe outcome and is consistent with the labeling relationship defined in Section 3.5, where default is treated as a specific severe form of enterprise distress.

The normalized outcome-severity score is defined as:

$$R_i = \begin{cases} 0, & no\ observed\ distress\ event, \\ 0.25, & significant\ operational\ deterioration, \\ 0.50, & financing\ failure, \\ 0.75, & severe\ liquidity\ crisis, \\ 1.00, & default\ event. \end{cases}$$

This score is constructed from future outcomes as a supervised target, not as an input feature. This separation alone does not rule out leakage through documents, retrieval, or sample timing.

The predicted continuous risk score $\hat{R}_i$ is obtained from the lifecycle-aware risk-level prediction head. Let $p_{i,l}$ denote the predicted probability of risk level l, and let $s_l \in \{0,0.25,0.50,0.75,1.00\}$ denote the corresponding normalized severity value. The predicted risk score is calculated as the probability-weighted expected severity:

© 2026 The Author(s). Licensed under CC BY 4.0

$$\hat{R}_i = \sum_{l=1}^{5} p_{i,l}\, s_l$$

The Root Mean Square Error (RMSE) is then calculated between the predicted continuous risk score and the ground-truth outcome severity:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (\hat{R}_i - R_i)^2}$$

Lower RMSE indicates closer agreement with the encoded outcome severity. It does not measure the calibration of distress probabilities or establish investment utility; its interpretation depends on the specified severity coding.

4.4 Results

To verify the effectiveness of R2D-LLM in enterprise distress prediction, the proposed framework is compared with several representative baseline models, including Logistic Regression (LR), Random Forest (RF), XGBoost, Long Short-Term Memory Network (LSTM), Graph Neural Network (GNN), Transformer, and FinGPT-based risk classification models. The comparison results are presented in Table 2.

Table 2Test-set performance comparison of R2D-LLM and seven baseline models for twelve-month private enterprise distress prediction.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
Logistic Regression	78.4	74.6	71.9	73.2	0.801
Random Forest	82.1	79.3	76.8	78.0	0.842
XGBoost	84.7	82.5	80.4	81.4	0.867
LSTM	86.2	83.8	82.7	83.2	0.881
GNN	87.5	85.6	84.3	84.9	0.895
Transformer	88.9	87.1	85.8	86.4	0.912
FinGPT-based Model	90.3	88.6	87.5	88.0	0.923
R2D-LLM (Ours)	92.8	91.4	89.9	90.6	0.936

Table 2 reports a consistent numerical advantage for R2D-LLM across all classification metrics. The complete model reaches 92.8% accuracy, 91.4% precision, 89.9% recall, a 90.6% F1-score, and an AUC of 0.936. Relative to the strongest baseline, the FinGPT-based model, these values represent gains of 2.5 percentage points in accuracy, 2.8 points in precision, 2.4 points in recall, and 2.6 points in F1-score, together with an AUC increase of 0.013. The simultaneous improvement in precision and recall suggests that the result is not produced only by favoring a more aggressive or more conservative classification threshold. The overall ranking from traditional machine-learning models to LSTM, GNN, Transformer, FinGPT, and R2D-LLM is consistent with the increasing use of nonlinear, temporal, relational, and domain-language information. However, the comparison alone does not isolate which component produces the gain; that question is addressed more directly by the ablation study. These are reported point estimates without uncertainty intervals; the AUC margin should not be interpreted as established statistical superiority.

While Table 2 evaluates the overall predictive performance of R2D-LLM against external baselines, Table 3 further examines whether the observed improvement depends on the individual modules proposed in the framework.

Table 3 reports model variants without RAG, risk-diffusion modeling, lifecycle information, or the stated optimization module. The optimization-module row is retained for author verification and is not interpreted as evidence of a predictive contribution.

© 2026 The Author(s). Licensed under CC BY 4.0

Table 3Reported component ablations on the chronological test set; the red optimization-module row requires author verification.

Model Variant	Accuracy (%)	F1-score (%)	AUC	RMSE
R2D-LLM without RAG	89.7	86.9	0.901	0.174
R2D-LLM without Risk Diffusion Transformer	90.5	87.8	0.912	0.169
R2D-LLM without Lifecycle Module	91.2	88.6	0.919	0.165
Complete R2D-LLM	**92.8**	**90.6**	**0.936**	**0.156**

Among the reported predictive ablations, removing RAG yields the largest decline: accuracy falls by 3.1 percentage points, F1-score by 3.7 points, and AUC by 0.035, while RMSE rises by 0.018. Removing the Risk Diffusion Transformer reduces F1-score by 2.8 points and AUC by 0.024; removing lifecycle information reduces them by 2.0 points and 0.017. These comparisons are consistent with useful retrieved, temporal, and lifecycle information, but do not establish extraction accuracy or causal propagation.

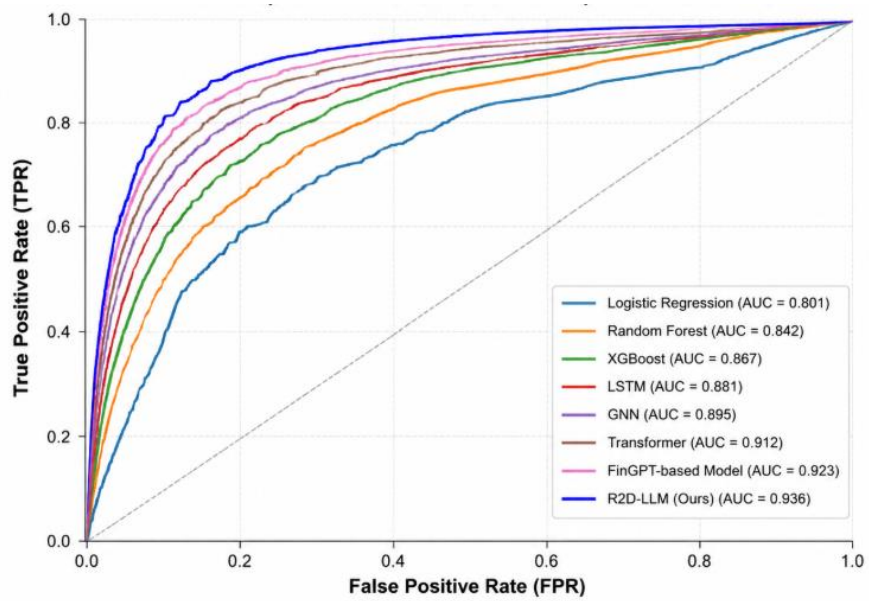


Figure 2Test-set ROC curves of R2D-LLM and seven baseline models for twelve-month private enterprise distress prediction. R2D-LLM achieves the highest AUC and maintains stronger discrimination across a broad range of classification thresholds.

Figure 2 complements Table 2 by showing discrimination across classification thresholds. R2D-LLM has the highest reported AUC, followed by the FinGPT-based model and the Transformer. The plotted ordering is not limited to the cutoff used for accuracy or F1-score. Exact sensitivity at low false-positive rates is not tabulated and should not be inferred numerically from the image. The figure therefore supports the reported ranking, but does not establish probability calibration, statistical superiority, or investment effectiveness.

Taken together, the comparative and ablation results indicate that the predictive advantage of R2D-LLM is associated with the joint use of retrieval-enhanced event representations, dynamic temporal modeling, and lifecycle information. The following discussion interprets these findings in relation to the paper's methodological motivation and practical application scope.

4.5 Discussion

Taken together, the comparison and ablation results indicate that R2D-LLM benefits from combining several forms of information that are usually modeled separately. Traditional machine-learning models provide a structured-data benchmark, while LSTM, GNN, and Transformer models introduce temporal, relational, and attention-based representations. The FinGPT-based baseline further incorporates financial-domain language understanding. Against this background, the additional improvement of R2D-LLM is consistent with the value of connecting retrieved domain knowledge, structured risk-event extraction, and

© 2026 The Author(s). Licensed under CC BY 4.0

dynamic event modeling. In particular, the gains in both precision and recall suggest that the model improves the separation between distressed and non-distressed enterprises rather than simply shifting the decision threshold. The reported RAG, diffusion, and lifecycle ablations are consistent with predictive relevance within the stated setup. They do not independently verify event-extraction accuracy or identify causal transmission. The unresolved optimization-module comparison is excluded from this interpretation.

The benchmark supports evaluation of enterprise risk ranking, subject to the cost of missed cases and false alarms. Results are point estimates without reported uncertainty, and no realized investment performance is evaluated. The downstream layer is therefore a decision-support concept rather than an established investment strategy. The above findings support the proposed framework as a dynamic distress-prediction approach, while also identifying several areas that require further validation before the downstream decision-support claims can be fully established.

5. Conclusions

This study proposes R2D-LLM, a retrieval-augmented large language model driven dynamic risk diffusion framework for private enterprise distress prediction. The framework is motivated by the limited financial disclosure, fragmented information, and rapidly changing risk conditions that characterize venture capital and private credit markets. It combines retrieval-enhanced risk-event extraction, dynamic risk-event graph construction, temporal risk diffusion modeling, and lifecycle-aware representation to transform heterogeneous enterprise documents and contextual variables into time-sensitive risk information.

The reported experiments show that R2D-LLM outperforms the evaluated machine-learning, deep-learning, Transformer, and financial-LLM baselines on the constructed multi-source dataset. The complete model achieves 92.8% accuracy, a 90.6% F1-score, and an AUC of 0.936. The ablation results further indicate that the RAG module and risk-diffusion component provide the largest measured contributions, while lifecycle information is associated with an additional improvement. The unresolved optimization-module comparison is not used to support these conclusions. The main contribution of the framework lies in connecting semantic risk understanding with temporal and relational event modeling. Rather than treating enterprise documents as static text features, R2D-LLM organizes extracted signals into structured events and models how these events accumulate and interact over time. This design provides a useful basis for early distress warning and dynamic enterprise risk monitoring. The downstream decision layer further illustrates how predicted risk states may be linked to candidate investment, holding, and exit actions. However, the current empirical evaluation primarily validates distress-prediction performance, and the effect of the decision layer on realized investment outcomes remains to be tested separately. The evidence is limited to the constructed predictive benchmark. Public documents may not capture private managerial information or unexpected shocks; event extraction has not been independently evaluated, and the reported results lack repeated-run uncertainty and probability-calibration analysis. Realized investment performance is also untested. These limitations bound the present claims rather than establish additional validated contributions.

© 2026 The Author(s). Licensed under CC BY 4.0

References

Brown, Tom, et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901.

Yang, Hongyang, Xiao-Yang Liu, and Christina Dan Wang. "Fingpt: Open-source financial large language models." arXiv preprint arXiv:2306.06031 (2023).

Wu, Shijie, et al. "Bloomberggpt: A large language model for finance." arXiv preprint arXiv:2303.17564 (2023).

Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." Advances in neural information processing systems 33 (2020): 9459-9474.

Gao, Yunfan, et al. "Retrieval-augmented generation for large language models: A survey." arXiv preprint arXiv:2312.10997 (2023).

Wang, Tangtang, Kaijie Zhang, and Kuangcong Liu. "A knowledge graph and deep learning-based semantic recommendation database system for advertisement retrieval and personalization." arXiv preprint arXiv:2601.00833 (2025).

Hawkes, Alan G. "Spectra of some self-exciting and mutually exciting point processes." Biometrika 58.1 (1971): 83-90.

Mei, Hongyuan, and Jason M. Eisner. "The neural hawkes process: A neurally self-modulating multivariate point process." Advances in neural information processing systems 30 (2017).

Zuo, Simiao, et al. "Transformer hawkes process." International conference on machine learning. PMLR, 2020.

Yoon, Kanghoon, et al. "Learning multivariate hawkes process via graph recurrent neural network." Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2023.

Mestiri, Sami. "Credit scoring using machine learning and deep Learning-Based models." Data Science in Finance and Economics 4.2 (2024): 236.

Mojdehi, Kosar Farajpour, Babak Amiri, and Amirali Haddadi. "A novel hybrid model for credit risk assessment of supply chain finance based on topological data analysis and graph neural network." IEEE Access 13 (2025): 13101-13127.

Jiang, Gaozhe, et al. "Investment advisory robotics 2.0: Leveraging deep neural networks for personalized financial guidance." 2025 International Joint Conference on Neural Networks (IJCNN). IEEE, 2025.

© 2026 The Author(s). Licensed under CC BY 4.0