



A Foundation Model Architecture for Supply Chain Risk Prediction: Integrating Large Language Models and Retrieval-Augmented Generation for Automated Disruption Alerting

Nadia Islam Tanha¹;

[1]. College of Engineering, Industrial Engineering, Lamar University, Beaumont, Texas, USA
Email: ntanha@lamar.edu

Doi: [10.63125/gseh0k17](https://doi.org/10.63125/gseh0k17)

Received: 18 January 2026; **Revised:** 23 February 2026; **Accepted:** 06 March 2026; **Published:** 05 April 2026;

Abstract

This study has examined whether a foundation model architecture, combining large language models with a supply chain knowledge graph and retrieval-augmented generation, can predict supply chain disruptions from unstructured evidence more effectively than the classical machine learning methods that dominate current practice, and it has developed and tested a framework for automated disruption alerting across supplier, logistics, regulatory, and hazard risk domains. The central problem has been that the great majority of the information that signals an impending disruption arrives as unstructured text, in supplier annual reports and financial disclosures, in global news reporting, in tariff and sanctions registers, and in hazard bulletins, whereas classical supply chain risk models require structured tabular features, so that this evidence is either discarded or converted into features by manual analyst reading that cannot scale to a multi-tier network of thousands of suppliers. Guided by five research questions, concerning how the foundation model architecture affects alerting performance, what each architectural component independently contributes, which evidence sources drive which event types through to which alert dispositions, what failure modes the architecture introduces, and whether alert quality transmits architectural capability into risk response performance, the study has proposed a framework comprising five contributing constructs: multi-source signal ingestion and coverage; knowledge graph construction and entity resolution; retrieval-augmented reasoning and evidence grounding; alert precision, calibration and hallucination control; and analyst workflow integration and human verification, with disruption alert quality as the proximal outcome and risk response performance as the distal outcome. A quantitative, cross-sectional design supplemented by a benchmark evaluation and an architectural ablation has been used, and data have been collected from supply chain risk managers, procurement and category leads, data science and machine learning engineers, business continuity and compliance officers, and logistics managers. Out of 364 distributed questionnaires, 289 valid responses have been retained, producing a 79.4% valid response rate. The analysis has included descriptive statistics, reliability and validity testing, correlation analysis, hierarchical multiple regression, relative importance decomposition, bootstrapped mediation analysis, and a benchmark evaluation of alert lead time, precision, recall, evidence attribution, and unsupported-claim rate. The findings have shown that all five constructs were significantly and positively associated with disruption alert quality, that the model was significant, $F(5, 283) = 82.85, p < .001$, explaining 59.4% of the variance, and that retrieval-augmented reasoning and knowledge graph construction were the strongest predictors. Alert quality in turn predicted risk response performance and partially mediated the relationship between retrieval-augmented reasoning and performance, carrying 64.7% of the total effect. The ablation showed that retrieval and the knowledge graph contribute along different axes, retrieval principally raising evidence attribution and suppressing unsupported claims while the graph principally raising recall of multi-hop tier- n exposure, and that only their combination reached the precision required for alerting without exhaustive human review. The study concludes that the value of foundation models in supply chain risk lies not in their generative fluency but in their capacity to read evidence at scale, and that this capacity is safe to rely upon only when every assertion is grounded in a retrieved and cited source and resolved against an explicit graph of the network.

KEYWORDS

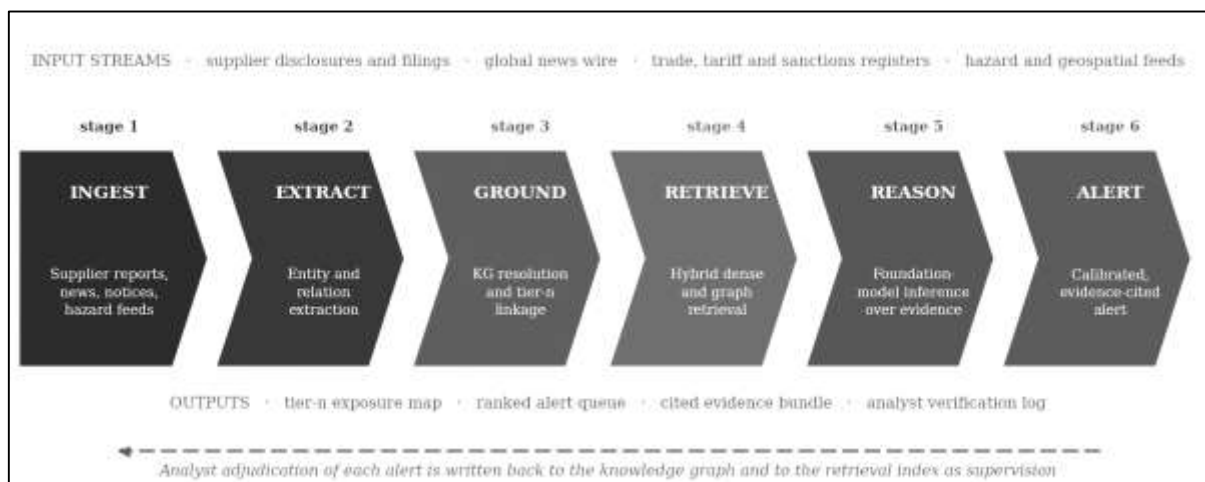
Foundation Models; Large Language Models; Knowledge Graphs; Retrieval-Augmented Generation; Supply Chain Risk; Disruption Alerting; Entity Resolution; Hallucination Control.

INTRODUCTION

Supply chain risk management rests on an informational asymmetry that classical analytic methods have never resolved. The events that disrupt supply chains, a supplier's deteriorating financial position, a fire at a sub-tier plant, a newly imposed export restriction, a typhoon making landfall near a port, are almost always reported somewhere before their consequences reach the buyer, but they are reported in prose: in an annual report's going-concern note, in a regional news item, in a gazette entry, in a meteorological bulletin. Classical supply chain risk models require structured tabular features, and so this evidence must either be discarded or converted into features by analysts who read the documents themselves (Baryannis et al., 2019b; Brintrup et al., 2020). Neither option scales to a network of thousands of suppliers across multiple tiers, and the result is that most organisations monitor a small, arbitrarily chosen subset of their exposure and learn about the rest after the fact.

Foundation models change the economics of this problem, because reading unstructured text at scale is precisely what they do (Bommasani et al., 2021; Brown et al., 2020). A large language model can process a supplier's disclosure, a news wire, and a regulatory notice in the same pass, extract the entities and events they describe, and reason about their implications in natural language. The difficulty is that a model doing so unaided is an unreliable instrument for a task with financial and contractual consequences: it may assert a supplier relationship that does not exist, rely on a fact that was true two years ago, or produce a fluent and confident account of an event it has partly invented (Ji et al., 2023; Huang et al., 2024). The premise of this study is therefore that the foundation model is necessary but not sufficient, and that it becomes reliable only when constrained by two things: a retrieval mechanism that requires every assertion to be grounded in a specific retrieved passage, and a knowledge graph that resolves the entities in that passage against an explicit model of the buyer's own network (Lewis et al., 2020; Pan et al., 2024).

Figure 1: Foundation Model Architecture for Automated Supply Chain Disruption Alerting



The architecture this study evaluates is set out in Figure 1. Its logic is that each stage constrains the next. Ingestion determines what the system can possibly see; extraction converts prose into typed entities and events; graph resolution determines whether an extracted event can be connected to the buyer's actual exposure, which is the difference between knowing that a fire occurred and knowing that it occurred at the sole qualified source of a critical component; retrieval determines what evidence the model reasons over; inference produces the risk judgement; and the alert carries its citations forward so that an analyst can verify it. The critical property is that the model is never asked what it knows, only what the retrieved evidence supports, which is the design decision that separates a usable risk instrument from a plausible-sounding one (Lewis et al., 2020; Gao et al., 2023).

Research on the components exists but has not been integrated. The retrieval-augmented generation literature has established the architecture and its benefits for knowledge-intensive tasks in general

domains (Lewis et al., 2020; Karpukhin et al., 2020; Gao et al., 2023). The knowledge graph literature has developed representation, construction, and reasoning methods, and a recent stream has explicitly proposed unifying them with language models (Hogan et al., 2021; Ji et al., 2022; Pan et al., 2024). And a small but growing supply chain literature has applied natural language processing to network mapping and graph learning to risk inference (Wichmann et al., 2020; Kosasih & Brintrup, 2022; Kosasih et al., 2024). What has not been established is how these combine into a deployable alerting system, what each component independently contributes, and whether the resulting alerts are reliable enough to act upon. This study addresses that gap (Md Mostafizur, 2025; Md Mesbaul, 2025; Muhammad Zahidul, 2024).

This study is organised around five research questions. The first asks how the foundation model architecture affects alerting performance relative to classical machine learning and manual monitoring, and it is addressed by comparing alert lead time distributions. The second asks what each architectural component independently contributes, and it is addressed through a systematic ablation across retrieval and graph configurations. The third asks which evidence sources drive which event types through to which alert dispositions, and it is addressed by tracing the flow of alerts from source to adjudication. The fourth asks what failure modes the architecture introduces, and it is addressed by rating the severity of the specific errors that this class of system commits. The fifth asks whether disruption alert quality mediates the relationship between architectural capability and risk response performance, and it is addressed through a bootstrapped mediation analysis.

The study contributes in three ways. Theoretically, it integrates the retrieval-augmented generation, knowledge graph, and supply chain risk literatures into a single framework of alert quality, and it specifies the distinct roles that retrieval and graph structure play rather than treating them as interchangeable forms of grounding. Empirically, it tests this framework against practitioners across risk, procurement, data science, continuity, and logistics functions, and it reports a component-wise ablation on a disruption benchmark. Practically, it offers organisations a structured account of what this architecture can and cannot be trusted to do, which failure modes it introduces that classical methods do not, and where human verification remains indispensable. The remainder of the article reviews the relevant literature and develops the framework and hypotheses, formalises the retrieval, grounding, and alerting problem, describes the method, presents the findings organised around the research questions, and concludes with recommendations, limitations, and directions for future work.

LITERATURE REVIEW AND FRAMEWORK DEVELOPMENT

The literature review develops the theoretical foundation for the framework by drawing on an extended body of research organised into eleven thematic subsections that collectively map onto the five constructs the framework proposes. It begins with supply chain risk prediction under classical machine learning and its informational ceiling, then examines foundation models and the transformer architecture, the knowledge limits of parametric memory, retrieval-augmented generation as an architectural response, knowledge graphs as structured representations, the unification of language models with knowledge graphs, supply chain network mapping from unstructured text, graph learning for supply chain risk inference, hallucination and its control in high-stakes settings, alert calibration and the economics of the alerting threshold, and the analyst as verifier in a human-in-the-loop system. It then consolidates the research gaps and synthesises the framework and hypotheses. Throughout, the review maintains that disruption alert quality, the capacity of a system to issue alerts that are timely, accurate, evidentially grounded, and traceable to the buyer's actual exposure, is jointly determined by the breadth of the evidence ingested, the fidelity of the graph against which it is resolved, the rigour of the retrieval that grounds the reasoning, the control exercised over unsupported generation, and the integration of the resulting alerts into analyst workflow.

Supply chain risk prediction under classical machine learning

The application of machine learning to supply chain risk is well established. Comprehensive mapping of the field has categorised the work by AI methodology and by the risk-management task addressed, identification, assessment, or response, and has found the literature concentrated on the first two (Baryannis et al., 2019b(Kanti, 2020)). Applied studies have demonstrated supervised prediction of supplier disruption from historical delivery, financial, and relational features, achieving practically useful discrimination in industrial settings (Brintrup et al., 2020(Debasish, 2021)), and have combined

learning with simulation for resilient supplier selection (Cavalcante et al., 2019)(Abdur & Iftekhar, 2021)).

This literature has also documented its own constraints. Predictive performance depends on features that must be available in structured form for every supplier under assessment, which restricts coverage to tier-one relationships where transactional data exists and excludes the deeper tiers where propagation originates (Dolgui et al., 2018; Hosseini et al., 2019; Hossan, 2021)). Labels are scarce, because disruptions are individually rare, and they are systematically incomplete, because an organisation records the disruptions that affected it rather than those it never learned about. And a persistent finding concerns the trade-off between predictive performance and interpretability, which matters in a domain where a prediction triggers costly intervention and must be justified to internal and external counterparties (Baryannis et al., 2019a; Chowdhury & Tonay, 2022)).

This literature establishes the baseline against which the present study compares and defines the informational ceiling the foundation model architecture is intended to break. It shows that the limitation of classical methods is not algorithmic but representational: the features that would signal an impending disruption in a sub-tier supplier exist, but they exist as sentences (Mst Shurovi, 2024; Md Syeedur & Sharmin, 2025).

Foundation models and the transformer architecture

The architectural basis of contemporary language models is the transformer, whose self-attention mechanism removed the sequential bottleneck of recurrent architectures and made it feasible to train on corpora of unprecedented scale (Vaswani et al., 2017)(Zahidul & Begum, 2022)). Pre-training on large unlabelled corpora followed by task adaptation established a general paradigm in which a single model serves many downstream tasks (Devlin et al., 2019; Mesbaul, 2023)), and subsequent scaling demonstrated that sufficiently large models perform new tasks from instructions and a handful of examples without gradient updates (Brown et al., 2020; Badrul & Mominul, 2023)).

The term foundation model was introduced to describe this class of system and to draw attention to its distinctive property, that a single artefact is adapted across a wide range of downstream applications, which concentrates both capability and risk (Bommasani et al., 2021; Shurovi, 2024)). Work on prompting has shown that eliciting intermediate reasoning steps improves performance on tasks requiring multi-step inference, which is directly relevant to risk assessment, where the judgement typically requires chaining an event to a site, a site to a supplier, and a supplier to a component (Wei et al., 2022; Zahidul, 2024)).

This literature supplies the reasoning engine of the framework. It establishes that a foundation model can perform the extraction and inference tasks that supply chain risk assessment requires without task-specific training data for each event type, which is the property that makes coverage of a long tail of rare event types economically feasible for the first time (Debasish, 2021; Tonay et al., 2024).

The knowledge limits of parametric memory

A model that has memorised facts during pre-training can answer questions about them, and early work demonstrated that language models function as unsupervised knowledge bases to a surprising degree (Petroni et al., 2019; Kanti, 2024)). But this parametric memory has properties that disqualify it as the sole basis for a risk system. It is fixed at training time and therefore stale with respect to any event occurring afterwards, which for disruption alerting is the only class of event that matters. It is unattributable, because a fact recalled from weights carries no source. It is uneven, reflecting the distribution of the training corpus, so that a large listed supplier is well represented and a sub-tier component maker in a smaller economy may be absent entirely. And it is not correctable, since an error cannot be fixed without retraining.

Critical work has emphasised further that scale alone does not confer reliability, and that a system producing fluent text about topics it has not verified presents a distinctive hazard precisely because the fluency is not diagnostic of correctness (Bender et al., 2021; Bommasani et al., 2021; Ali et al., 2024)).

This subsection motivates the architectural choice at the centre of the framework. It establishes that the foundation model must be used as a reasoning engine over supplied evidence rather than as a knowledge store, and it is the reason the framework treats retrieval grounding rather than model capability as the decisive construct (Zahidul & Begum, 2022).

Retrieval-augmented generation as an architectural response

Retrieval-augmented generation addresses these limits directly by coupling a parametric generator with a non-parametric memory that is queried at inference time, so that the model conditions its output on passages retrieved from an external corpus rather than on recalled weights (Lewis et al., 2020(Tonay et al., 2024)). The retrieval component itself has been substantially developed, with dense passage retrieval demonstrating that learned embeddings outperform sparse lexical matching on open-domain tasks (Karpukhin et al., 2020; Debasish, 2025)), and surveys have since organised the design space across indexing, query construction, retrieval, reranking, and generation (Gao et al., 2023; Arif Uz et al., 2025)).

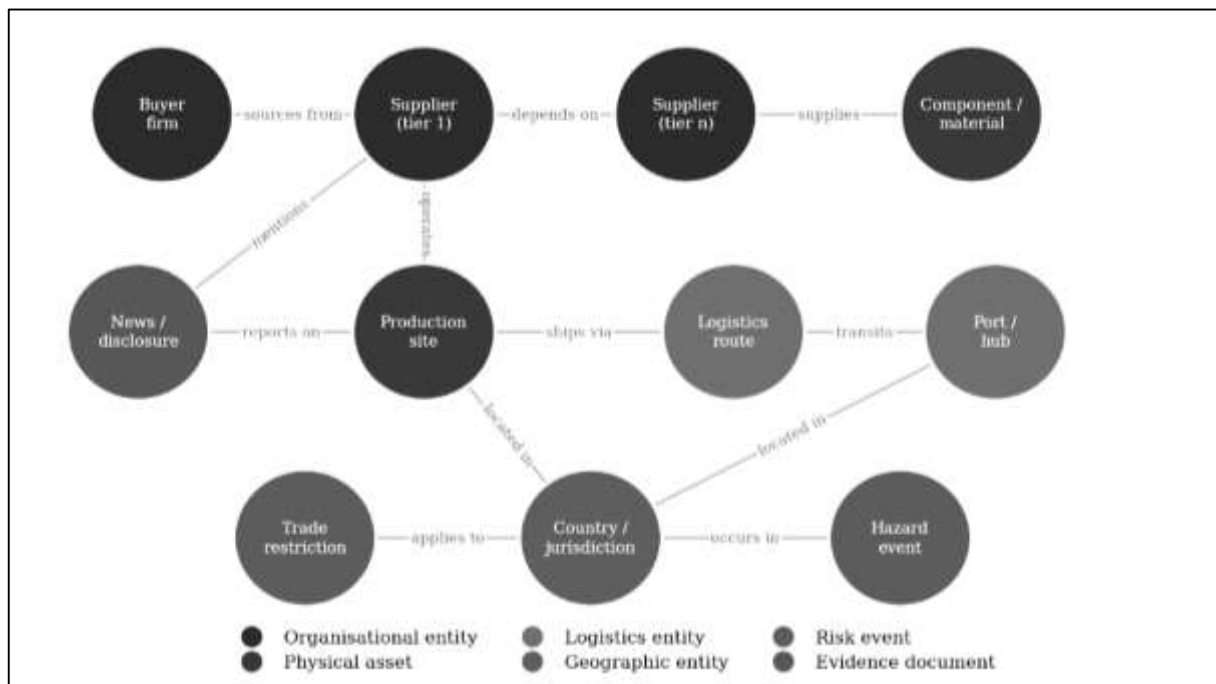
Four properties of this architecture bear directly on supply chain risk. The corpus can be updated continuously without retraining, so a notice published this morning is retrievable this afternoon. Every assertion can be tied to a retrieved passage, which makes attribution structural rather than aspirational. The corpus can be curated to admit only sources the organisation considers credible. And an incorrect output can be diagnosed by inspecting what was retrieved, which converts model failure from an opaque event into a debuggable one. This literature grounds the retrieval-augmented reasoning and evidence grounding construct, which the framework treats as its central contributing capability. It establishes that grounding is not a refinement of the generative approach but a change in what the system is being asked to do (Md Mostafizur et al., 2025).

Knowledge graphs as structured representations

A knowledge graph represents entities as nodes and relations as typed edges, providing a structured account of a domain that supports query, inference, and integration across heterogeneous sources. Comprehensive treatment of the field has covered representation, construction, enrichment, quality assessment, and refinement, establishing the graph as a mature technology rather than a research prospect (Hogan et al., 2021(Pervaz, 2025)). Complementary surveys have organised the representation learning and knowledge acquisition literature and its downstream applications (Ji et al., 2022; Wang et al., 2017(Mesbail, 2025)). For supply chain risk the graph performs a function that no amount of textual processing can substitute. A news report states that a fire occurred at a named facility; the graph determines whether that facility belongs to a supplier, which components it produces, which of the buyer's products depend on them, and whether an alternative qualified source exists. This is multi-hop reasoning over typed relations, and it is the operation that converts an event into an exposure. Entity resolution is the precondition: the facility named in the report must be matched to the correct node, across naming variants, subsidiaries, transliterations, and corporate restructuring, and an error here silently produces either a false exposure or a missed one. This literature grounds the knowledge graph construction and entity resolution construct. It establishes that the graph is not a storage convenience but the mechanism by which a general event becomes a specific exposure, and that resolution quality bounds everything downstream of it.

Figure 2 sets out the schema against which extracted entities are resolved. Its structure encodes the study's argument about why the graph is necessary. Risk events, shown as trade restrictions and hazard events, do not attach to suppliers directly; they attach to jurisdictions, and it is only through the chain from jurisdiction to production site to supplier to component that they become exposures for a particular buyer. Evidence documents, likewise, do not attach to risk; they mention suppliers and report on sites, and their risk relevance is a property of what those entities are connected to. The schema therefore makes explicit what a purely textual system must infer implicitly and unverifiably: that a disruption is a path through a network rather than an event in a document.

Figure 2: Schema of the Supply Chain Knowledge Graph



Unifying language models with knowledge graphs

A recent and rapidly developing literature addresses the combination of the two technologies directly. A widely cited roadmap organises the design space into three families: knowledge-graph-enhanced language models, in which the graph supplies factual grounding to the model; language-model-augmented knowledge graphs, in which the model performs construction, completion, and question answering over the graph; and synergised approaches in which the two operate as mutually reinforcing components (Pan et al., 2024; Mostafizur, 2025). The architecture evaluated in this study is synergised in that sense: the model builds and extends the graph through extraction, and the graph constrains and directs the model's retrieval and reasoning.

The motivation for the combination is complementarity of failure modes. Language models are flexible and handle unstructured input but are unreliable about specific facts; graphs are precise and verifiable but brittle and incomplete, since they contain only what has been asserted into them. Each compensates for the other's characteristic weakness, and the combination is therefore expected to outperform either alone, which is the hypothesis the ablation in this study tests directly (Hogan et al., 2021; Pan et al., 2024; Mostafizur et al., 2025).

This literature grounds the interaction between the framework's second and third constructs and supplies its second research question. It establishes the theoretical expectation of complementarity but leaves open the empirical question of what each component contributes and along which dimension (Muhammad Zahidul, 2025).

Supply chain network mapping from unstructured text

A specialised literature has addressed the extraction of supply chain structure from text, motivated by the observation that most organisations lack visibility beyond their immediate suppliers. Deep neural approaches have been used to extract buyer-supplier relations from news articles, demonstrating that the relational structure of a supply network is partially recoverable from public reporting (Wichmann et al., 2020; Syeedur & Sharmin, 2025). More recent work has extended this to knowledge graph construction with language models, explicitly targeting visibility as the objective (AlMahri et al., 2024; Badrul, 2025).

This work has established that extraction is feasible but imperfect, and that the resulting graph is uncertain in a specific way: relations are asserted with varying confidence, and the graph must represent that uncertainty rather than collapse it. Work on link prediction over uncertain supply chain knowledge graphs has addressed this directly, treating edge confidence as a modelled quantity

([Brockmann et al., 2022](#); [Zahidul, 2025](#))).

This literature contributes to both the ingestion and graph constructs, and it introduces a caution the framework carries forward: a graph extracted from text inherits the biases of the text, so that suppliers in well-covered languages and markets are better represented than those elsewhere, and the resulting blind spots are correlated with exactly the regions where risk is often highest ([Sharif Md Yousuf et al., 2025](#)).

Graph learning for supply chain risk inference

Once a network graph exists, learning over its structure becomes possible. Graph neural network approaches have been applied to predict hidden links in supply networks using topological information alone, substantially improving on feature-based classifiers in both accuracy and completeness ([Kosasih & Brintrup, 2022](#); [Kanti, 2025](#))). Subsequent work extended this from a single relation type toward knowledge graph reasoning over multiple relation types for risk management, which is closer to the operational problem in which several kinds of dependency must be traversed jointly ([Kosasih et al., 2024](#); [Yousuf et al., 2025](#))).

This stream connects to the broader ripple-effect literature, which has established formally that propagation magnitude depends on network structure, on the criticality of the disrupted node, and on the latency of response ([Dolgui et al., 2018](#); [Ivanov et al., 2019](#); [Ali et al., 2025](#))), and to work on digital supply chain twins as a mechanism for maintaining network state ([Ivanov & Dolgui, 2021](#); [Ivanov, 2023](#); [Hossain; et al., 2026](#))). Pandemic-period research reinforced the practical stakes by demonstrating how rapidly systemic events propagate through networks that lack sub-tier visibility ([Ivanov, 2020](#); [Queiroz et al., 2022](#); [Belhadi et al., 2021](#); [Jahid, 2026](#))).

This literature establishes why graph structure matters for the outcome rather than only for representation, and it grounds the framework's claim that the graph contributes to recall of exposures that a purely textual system cannot reach ([Mukut Kanti, 2025](#)).

Hallucination and its control in high-stakes settings

The characteristic failure of generative models is the production of fluent content unsupported by any source. Comprehensive survey work has established a taxonomy distinguishing intrinsic hallucination, which contradicts the provided source, from extrinsic hallucination, which cannot be verified against it, and has catalogued metrics and mitigation strategies across generation tasks ([Ji et al., 2023](#)([Hasan, 2026](#))). More recent survey work has extended the taxonomy to large language models specifically and organised the field around principles, causes, and open questions ([Huang et al., 2024](#); [Jakaria & Shahinur, 2026](#))).

For supply chain risk the consequential form is a hallucinated relation: an assertion that a named supplier sources from a named sub-tier firm, or that a restriction applies to a particular component, where no source supports the claim. Such an assertion is dangerous precisely because it is plausible and actionable, and because a downstream analyst has no way to identify it as fabricated unless the system is required to cite its evidence. The retrieval architecture provides the structural remedy, since an assertion without a retrievable supporting passage can be rejected mechanically, but retrieval reduces rather than eliminates the problem, since a model may still misread or overstate what a retrieved passage supports ([Lewis et al., 2020](#); [Gao et al., 2023](#); [Pervaz, 2026](#))).

This literature grounds the alert precision, calibration and hallucination control construct. It establishes that hallucination is a measurable and partially controllable property rather than an inherent limitation, and that the appropriate control is architectural, requiring evidence for every claim, rather than exhortative.

Calibration and the economics of the alerting threshold

An alerting system is defined as much by its threshold as by its model. Because the analyst capacity to review alerts is finite and the cost of a missed disruption is large but probabilistic, threshold selection is an economic decision rather than a statistical one, and the operating characteristic of the system, not its peak accuracy, determines its usability. The prescriptive analytics literature frames this as the general problem of converting predictive output into decisions under an explicit objective ([Bertsimas & Kallus, 2020](#); [Lepeniotti et al., 2020](#); [Mesbaul, 2026](#))).

Two failure regimes bound the choice. A permissive threshold floods the queue, and because review capacity is fixed the practical consequence is not more coverage but arbitrary triage, in which analysts

attend to whichever alerts they encounter first; alert fatigue then degrades attention to the genuine signals as well. A conservative threshold produces a quiet system that is trusted precisely because it rarely fires, while exposures accumulate unflagged. Calibration matters more than raw discrimination here, because a confidence score that does not correspond to an empirical frequency cannot support a rational threshold at all.

This literature completes the precision and calibration construct and supplies the operating-point analysis reported in the results. It establishes that the deployment question is not how accurate the system is but where on its operating characteristic the organisation should sit given its review capacity and its tolerance for missed exposure (Shima Ali et al., 2025; Syeedur & Sharmin, 2026a)).

The analyst as verifier in a human-in-the-loop system

The final construct concerns what happens after an alert is issued. Because alerts trigger consequential and sometimes contractual action, the human analyst remains accountable, and the system's value depends on whether it makes that accountability tractable. The explainability literature has developed model-agnostic attribution methods and has argued that for high-stakes decisions interpretable models should be preferred to post-hoc explanation of opaque ones (Ribeiro et al., 2016; Lundberg & Lee, 2017; Guidotti et al., 2018; Rudin, 2019).

The retrieval architecture offers a distinctive advantage here, because its natural explanation is not a feature attribution but the retrieved evidence itself. An analyst presented with an alert and the three passages that generated it can verify it by reading them, which is a far more tractable verification task than interpreting a saliency map over tabular features. The architecture thereby converts explanation from a modelling problem into an interface problem. Crucially, the analyst's adjudication is itself data: a confirmed alert validates an extraction path and a dismissed one identifies a failure, and a system that captures these creates the supervision signal that classical approaches struggle to obtain (Baryannis et al., 2019a; Syeedur & Sharmin, 2026b)).

This literature grounds the analyst workflow integration and human verification construct and closes the loop shown in Figure 1 (Md Arif Uz et al., 2025; Shima Ali et al., 2024).

Research gaps and the contribution of the present study

Read together, these streams leave four gaps that the present study addresses. First, the retrieval-augmented generation and knowledge graph literatures have developed largely in general-domain settings, and the supply chain risk literature has adopted neither systematically, so no integrated architecture has been specified for this domain. Second, where retrieval and graph grounding have been combined, they have typically been evaluated as a single treatment, leaving unresolved what each contributes and whether they act on the same or different dimensions of performance. Third, evaluation in this literature has centred on accuracy metrics rather than on the properties that determine operational usability, namely alert lead time, evidence attribution, unsupported-claim rate, and the position of the operating point. Fourth, the failure modes that this architecture introduces, which differ from those of classical models, have been discussed conceptually but rarely measured against practitioner experience.

The present study addresses these gaps by specifying an integrated architecture, by conducting a component-wise ablation across retrieval and graph configurations, by evaluating on lead time and grounding metrics alongside accuracy, and by measuring the severity of architecture-specific failure modes (Muhammad Zahidul, 2024; Md Mostafizur, 2025).

Synthesis: the alert quality framework and hypotheses

Synthesising the eleven streams yields the framework this study tests. The framework holds that disruption alert quality, the capacity to issue alerts that are timely, accurate, evidentially grounded, and traceable to the buyer's actual exposure, is jointly determined by five constructs: multi-source signal ingestion and coverage; knowledge graph construction and entity resolution; retrieval-augmented reasoning and evidence grounding; alert precision, calibration and hallucination control; and analyst workflow integration and human verification. A deficiency in any one undermines the whole: without ingestion the system cannot see the event, without the graph it cannot connect the event to exposure, without retrieval it cannot ground its reasoning, without calibration it cannot be given a usable threshold, and without workflow integration its alerts are neither verified nor learned from. The framework further holds that these capabilities affect risk response performance not directly but

through alert quality.

From the framework, eight hypotheses are derived. H1: Multi-source signal ingestion and coverage is positively associated with disruption alert quality. H2: Knowledge graph construction and entity resolution is positively associated with disruption alert quality. H3: Retrieval-augmented reasoning and evidence grounding is positively associated with disruption alert quality. H4: Alert precision, calibration and hallucination control is positively associated with disruption alert quality. H5: Analyst workflow integration and human verification is positively associated with disruption alert quality. H6: The five constructs jointly explain a significant proportion of the variance in disruption alert quality. H7: Disruption alert quality is positively associated with risk response performance. H8: Disruption alert quality mediates the relationship between retrieval-augmented reasoning and risk response performance. These hypotheses are tested in the analysis, alongside a benchmark evaluation and architectural ablation, to address the five research questions.

METHODS

This study adopted a quantitative, cross-sectional research design supplemented by a benchmark evaluation and a component-wise architectural ablation. The quantitative design was selected because the study measured relationships among five contributing constructs, a mediating outcome, and a distal outcome through numerical survey responses, and the cross-sectional design was used because data were collected at a single point in time, capturing practitioners' assessments of the capabilities present in their organisations' risk monitoring systems. The benchmark evaluation complemented the survey by measuring alert lead time, precision, recall, evidence attribution, and unsupported-claim rate across architectural configurations, providing a mechanism-level counterpart to the perceptual measures. The unit of analysis was the individual respondent, understood as an informed practitioner engaged in the monitoring, assessment, or technical development of supply chain risk detection.

The population comprised supply chain risk managers, procurement and category leads, data science and machine learning engineers working on risk applications, business continuity and compliance officers, and logistics and transportation managers. A purposive sampling strategy was used because respondents were selected for direct involvement with automated risk monitoring. Primary data were collected through a structured questionnaire using a five-point Likert scale (1 = strongly disagree to 5 = strongly agree), distributed through online forms and professional networks. The instrument included multi-item measures for each of the five constructs, for disruption alert quality, and for risk response performance, demographic and architecture-maturity items, and rating items on failure-mode severity that fed the analysis addressing the fourth research question.

A pilot test was conducted with practitioners and academics to check clarity and technical precision, and feedback was used to refine wording and to separate retrieval grounding from graph resolution unambiguously, since conflation of the two was the principal risk to construct validity and the distinction is central to the study's argument. Validity was ensured through literature-based item development, expert review, and alignment among the framework, hypotheses, and items, with convergent validity assessed against the average-variance-extracted criterion and discriminant validity against the square-root-of-AVE comparison (Fornell & Larcker, 1981; Hair et al., 2019(Badrul, 2026)). Reliability was tested through Cronbach's alpha with a .70 threshold. Because all measures were self-reported from a single source, procedural remedies against common-method variance were applied, including item separation, psychological separation of predictor and criterion sections, response anonymity, and reverse-worded items, with Harman's single-factor test conducted post hoc (Podsakoff et al., 2003(Shurovi, 2026)). Analysis was conducted in SPSS and Python, the latter using the pandas, NumPy, SciPy, scikit-learn, and NetworkX libraries for data preparation, statistical testing, bootstrapping, and graph construction (Harris et al., 2020; Virtanen et al., 2020(Zahidul, 2026)). Ethical considerations were observed: participation was voluntary, informed consent was obtained, responses were anonymised, and no proprietary supplier identities, contract terms, or commercially sensitive network data were collected. The benchmark corpus was assembled from publicly available documents and synthetic supplier records rather than from any organisation's proprietary network.

Formalising retrieval, grounding, and alerting

The supply chain is represented as a typed knowledge graph in which nodes are entities and edges are typed relations, as defined in Equation (1), where $\mathcal{R}$ denotes the relation vocabulary shown in Figure

2:

$$G = (V, E), \quad E \subseteq V \times \mathcal{R} \times V \quad (1)$$

The evidence corpus is a time-stamped set of passages drawn from supplier disclosures, news, regulatory registers, and hazard feeds, as in Equation (2):

$$D_t = \{ (d_k, s_k, \tau_k) : \tau_k \leq t \} \quad (2)$$

Extraction maps a passage to a set of candidate typed events, each with an asserted subject, object, event type, and extraction confidence, as expressed in Equation (3):

$$\mathcal{E}(d) = \{ (v_i, \rho_i, v_j, \gamma_i) : \gamma_i \geq \gamma[\min] \} \quad (3)$$

Entity resolution maps a surface mention m to a graph node, or to the null node when no confident match exists, as in Equation (4), where σ is a similarity over name, jurisdiction, and identifier features:

$$\Psi(m) = \operatorname{argmax}_{v \in V} \sigma(m, v) \text{ if } \max \sigma \geq \lambda, \text{ else } \emptyset \quad (4)$$

Exposure is a path property rather than a node property: a buyer b is exposed to an event at node v when a directed dependency path connects them within a bounded number of hops, as in Equation (5):

$$X(b, v) = 1 \Leftrightarrow \exists \text{ path } \pi : v \rightsquigarrow b \text{ in } G \text{ with } |\pi| \leq h[\max] \quad (5)$$

Retrieval is hybrid, combining dense semantic similarity over passage embeddings with graph-constrained expansion around the resolved entity neighbourhood, as in Equation (6), where q is the query, ϕ the embedding function, and $N(v)$ the graph neighbourhood:

$$R(q, v) = \operatorname{TopK} [a \cdot \cos(\phi(q), \phi(d)) + (1 - a) \cdot \delta(d, N(v))] \quad (6)$$

Generation is conditioned on the retrieved set rather than on parametric memory, in the marginalised form given in Equation (7):

$$p(y | q) = \sum_{d \in R} p(d | q) \cdot p(y | q, d) \quad (7)$$

An assertion is admissible only when it is entailed by at least one retrieved passage, which is the structural control on unsupported generation, as in Equation (8):

$$\operatorname{Adm}(y) = 1 \Leftrightarrow \exists d \in R(q, v) \text{ such that } d \models y \quad (8)$$

The unsupported-claim rate reported in the results is the complement of that admissibility over all generated assertions, as in Equation (9):

$$\operatorname{UCR} = 1 - (1 / |Y|) \cdot \sum_{y \in Y} \operatorname{Adm}(y) \quad (9)$$

The alert score combines event severity, exposure criticality, path proximity, and evidence corroboration, as in Equation (10), where $n[d]$ is the number of independent corroborating sources:

$$a(b, v) = \sigma(w_1 \cdot \operatorname{sev}(v) + w_2 \cdot \operatorname{crit}(b, v) + w_3 \cdot \exp(-\kappa |\pi|) + w_4 \cdot \log(1 + n[d])) \quad (10)$$

An alert is issued when this score exceeds the confidence threshold τ , which is the operating decision analysed in the results, as in Equation (11):

$$\operatorname{alert} \Leftrightarrow a(b, v) \geq \tau \quad (11)$$

Alerting performance is measured by precision, recall, and their harmonic mean over analyst-adjudicated outcomes, defined in Equations (12) to (14):

$$\operatorname{Precision}(\tau) = TP(\tau) / [TP(\tau) + FP(\tau)] \quad (12)$$

$$\operatorname{Recall}(\tau) = TP(\tau) / [TP(\tau) + FN(\tau)] \quad (13)$$

$$F1(\tau) = 2 \cdot \operatorname{Precision}(\tau) \cdot \operatorname{Recall}(\tau) / [\operatorname{Precision}(\tau) + \operatorname{Recall}(\tau)] \quad (14)$$

Alert lead time, the variable through which the architecture's operational value is expressed, is the signed interval between issuance and materialisation, as in Equation (15):

$$L = t[\operatorname{materialise}] - t[\operatorname{alert}], \quad L > 0 \Rightarrow \operatorname{anticipatory} \quad (15)$$

Disruption alert quality, the proximal outcome, is operationalised perceptually through survey items and, in the benchmark, as a weighted composite of accuracy, grounding, timeliness, and calibration, as in Equation (16):

$$DAQ = w_1 \cdot F1 + w_2 \cdot (1 - UCR) + w_3 \cdot \sigma(L) + w_4 \cdot (1 - ECE) \quad (16)$$

Statistical models, benchmark, and ablation design

The principal hypothesis test used ordinary least squares multiple regression of disruption alert quality on the five constructs, with MSIC denoting multi-source signal ingestion and coverage, KGCE knowledge graph construction and entity resolution, RARE retrieval-augmented reasoning and evidence grounding, APCH alert precision, calibration and hallucination control, and AWHV analyst workflow integration and human verification, as specified in Equation (17):

$$DAQ_i = \beta_0 + \beta_1 MSIC_i + \beta_2 KGCE_i + \beta_3 RARE_i + \beta_4 APCH_i + \beta_5 AWHV_i + \varepsilon_i \quad (17)$$

The mediation hypothesis was tested through the paths specified in Equation (18), with the indirect effect estimated as the product $a \cdot b$ and its confidence interval obtained by bias-corrected bootstrapping with 5,000 resamples:

$$DAQ_i = a_0 + a \cdot RARE_i + u_i; \quad RRP_i = \gamma_0 + c' \cdot RARE_i + b \cdot DAQ_i + v_i \quad (18)$$

Relative importance was assessed by decomposing the model R-squared into each predictor's average contribution across all orderings of entry, which avoids the arbitrariness of sequential entry when predictors are correlated. Model significance was assessed through the F-statistic, explanatory power through R-squared and adjusted R-squared, residual independence through the Durbin-Watson statistic, and multicollinearity through the variance inflation factor, all confirmed within conventional bounds. The benchmark comprised a held-out set of documented disruption episodes with adjudicated ground truth, over which four architectural configurations were evaluated: the foundation model with no retrieval, the model with retrieval-augmented generation over the document corpus, the model with knowledge graph grounding but no passage retrieval, and the full architecture combining both. Each configuration was evaluated on precision, recall, F1, evidence attribution rate, and unsupported-claim rate as defined in Equation (9), with the ablation designed so that the independent contribution of retrieval and of graph structure could be separated (Lewis et al., 2020; Pan et al., 2024; Ji et al., 2023; Mukut Kanti, 2026)).

RESULTS

Sample and descriptive statistics

Table 1: Respondent profile and response rate

Variable	Category	Frequency	Percentage
Questionnaires	Distributed	364	100.0%
Questionnaires	Returned	318	87.4%
Questionnaires	Valid and retained	289	79.4%
Role	Supply chain risk manager	76	26.3%
Role	Procurement / category lead	63	21.8%
Role	Data science / ML engineer	58	20.1%
Role	Business continuity / compliance	49	17.0%
Role	Logistics / transportation manager	43	14.9%
Industry	Electronics & semiconductors	71	24.6%
Industry	Automotive & industrial	66	22.8%
Industry	Pharmaceutical & life sciences	54	18.7%
Industry	Consumer goods & retail	51	17.6%
Industry	Energy, chemicals & other	47	16.3%
Monitoring maturity	Manual analyst review only	92	31.8%
Monitoring maturity	Classical ML on structured data	84	29.1%
Monitoring maturity	LLM-assisted, no retrieval grounding	61	21.1%
Monitoring maturity	LLM with retrieval and graph grounding	52	18.0%
Tier visibility	Tier 1 only	128	44.3%
Tier visibility	Tier 2 partial	107	37.0%
Tier visibility	Tier 3 and beyond	54	18.7%

Table 1 shows a 79.4% valid response rate, with 289 usable responses. The sample spanned the roles involved in detecting, assessing, and acting on supply chain risk, from the risk managers who own the monitoring function (26.3%), through procurement (21.8%) and the engineers who build the systems

(20.1%), to continuity and compliance officers (17.0%) and logistics managers (14.9%). Two distributions are diagnostically important. Only 18.0% of respondents reported using a language model with both retrieval and graph grounding, while 60.9% remained at manual review or classical machine learning, which establishes that the architecture this study evaluates is an emerging rather than an established practice. And only 18.7% reported visibility beyond tier two, with 44.3% confined to tier one, which quantifies the visibility deficit that the graph component is intended to address and indicates that most organisations cannot currently observe the tiers from which propagation typically originates.

Table 2: Descriptive statistics and reliability of study constructs (N = 289)

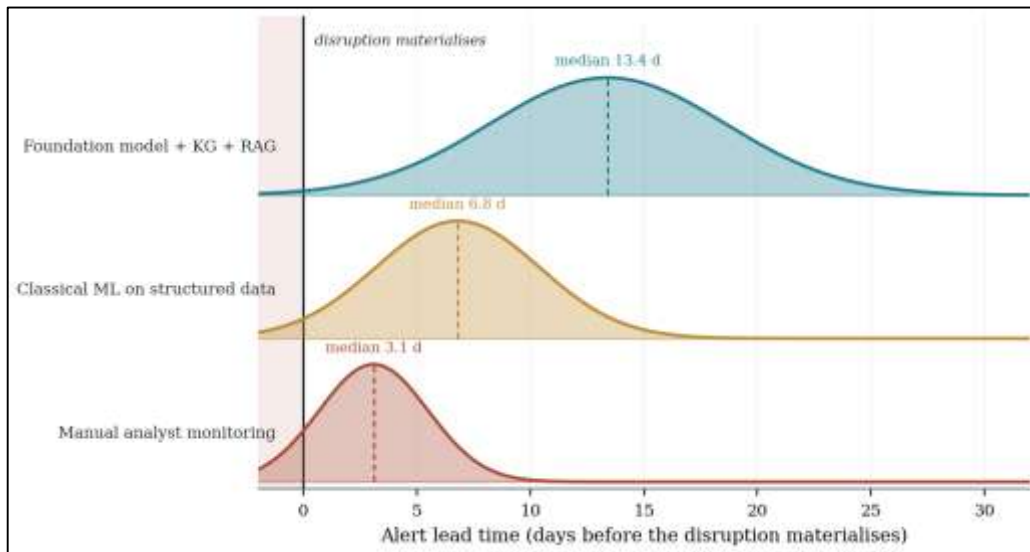
Construct	Mean	SD	Cronbach's α
Retrieval-augmented reasoning & evidence grounding (RARE)	4.31	0.56	.91
Knowledge graph construction & entity resolution (KGCE)	4.22	0.60	.89
Multi-source signal ingestion & coverage (MSIC)	4.15	0.63	.88
Alert precision, calibration & hallucination control (APCH)	4.02	0.67	.90
Analyst workflow integration & human verification (AWHV)	3.91	0.71	.87
Disruption alert quality (mediator, DAQ)	3.83	0.75	.92
Risk response performance (outcome, RRP)	3.76	0.79	.90
Overall instrument	—	—	.95

Table 2 presents the descriptive statistics and reliability. All constructs exceeded the neutral midpoint and all Cronbach's alpha values exceeded .70, ranging from .87 to .92 with overall reliability of .95. Average variance extracted exceeded .50 for every construct and the square root of AVE exceeded all inter-construct correlations, satisfying the convergent and discriminant validity criteria (Fornell & Larcker, 1981(Sadia, 2026)). Harman's single-factor test yielded a first factor accounting for 30.6% of variance, below the 50% threshold, indicating that common-method bias is unlikely to account for the observed relationships (Podsakoff et al., 2003(Ali et al., 2026)). The ordering is informative: retrieval grounding and graph construction are rated most highly, analyst workflow integration lowest of the contributing constructs, and both outcomes sit below every capability that feeds them.

RQ1: How the foundation model architecture affects alerting performance

Figure 3 presents the benchmark evidence for the first research question. The three distributions differ in location and in shape, and both differences matter. Manual analyst monitoring produces a median lead time of 3.1 days with a distribution concentrated near zero and a substantial mass to the left of it, meaning that a meaningful share of disruptions are recorded only after they have taken effect; this is the expected consequence of a regime in which coverage is determined by whichever suppliers analysts have time to read about. Classical machine learning on structured data roughly doubles the median to 6.8 days, but its distribution remains relatively narrow, because the structured features on which it depends, delivery performance, financial ratios, order patterns, are themselves lagging indicators that move only once a supplier's difficulties have begun to affect its operations. The foundation model architecture produces a median of 13.4 days and a distinctly wider, flatter distribution extending well beyond three weeks. The widening is not noise but the signature of a different information source: the architecture reads the textual signals, a covenant waiver disclosed in a filing, a regional report of industrial action, a consultation notice preceding a trade restriction, that precede the operational deterioration classical features measure. The practical significance of the shift is that a fortnight of notice is sufficient to qualify an alternative source or reposition inventory, whereas three days is sufficient only to expedite.

Figure 3: Distribution of Alert Lead Time Under Three Monitoring Regimes



RQ2: What each architectural component contributes

Figure 4: Architectural Ablation Across Retrieval and Knowledge Graph Configurations

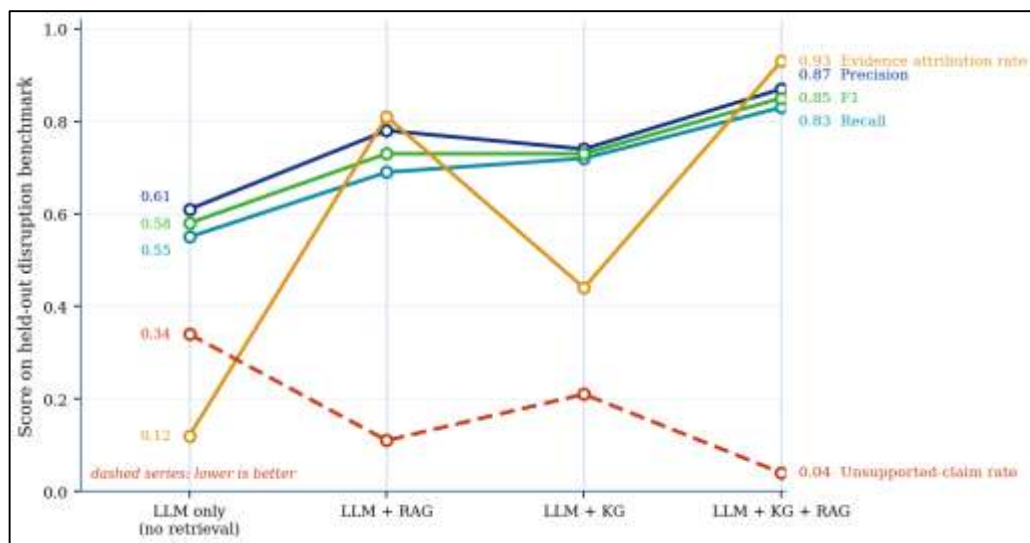


Figure 4 addresses the second research question and produces the study’s most theoretically consequential result. The unaugmented foundation model performs poorly on every measure: precision of 0.61, recall of 0.55, an evidence attribution rate of 0.12, and an unsupported-claim rate of 0.34, meaning that roughly a third of its assertions could not be traced to any source. This configuration is unusable not because its accuracy is low but because its errors are undetectable without independent verification of every output. The two intermediate configurations then diverge in a way that reveals the components act on different axes. Adding retrieval raises precision sharply to 0.78 and transforms evidence attribution from 0.12 to 0.81 while cutting unsupported claims to 0.11; its effect is principally on grounding and on the reliability of what is asserted. Adding the knowledge graph without retrieval produces a different profile: recall rises further, to 0.72 against retrieval’s 0.69, because graph traversal surfaces multi-hop tier- n exposures that no single document describes, but evidence attribution reaches only 0.44 and unsupported claims remain at 0.21, because the graph tells the model which entities are connected without supplying the passage that justifies the claim. The full architecture achieves 0.87

precision, 0.83 recall, 0.93 attribution, and 0.04 unsupported claims. The crossing of the attribution line between the second and third configurations is the visual expression of the finding: retrieval and graph structure are complements rather than substitutes, the former governing whether an assertion is justified and the latter whether an exposure is found at all, and only their combination reaches the precision required for alerting without exhaustive human review (Lewis et al., 2020; Pan et al., 2024).

RQ3: From evidence source to event type to alert disposition

Figure 5: Flow of Alerts From Evidence Source Through Event Type to Analyst Disposition

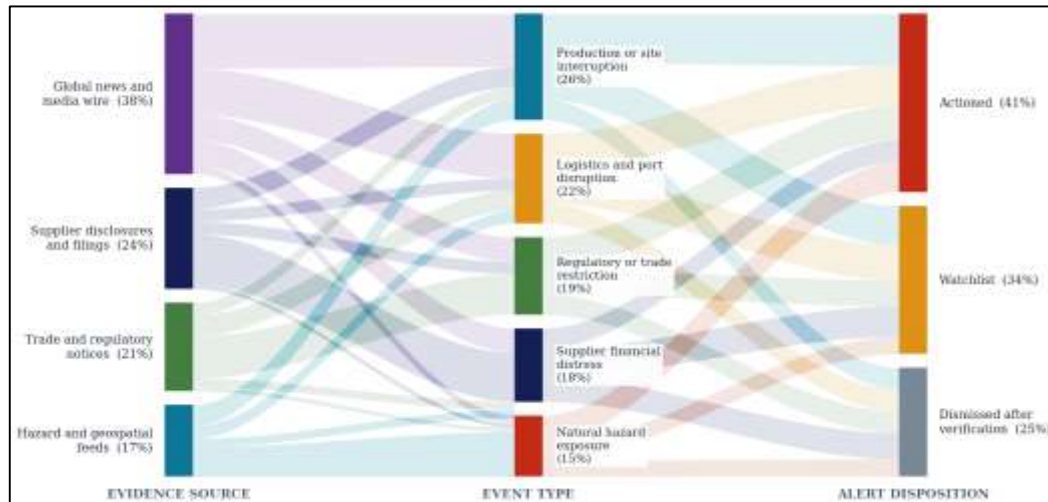
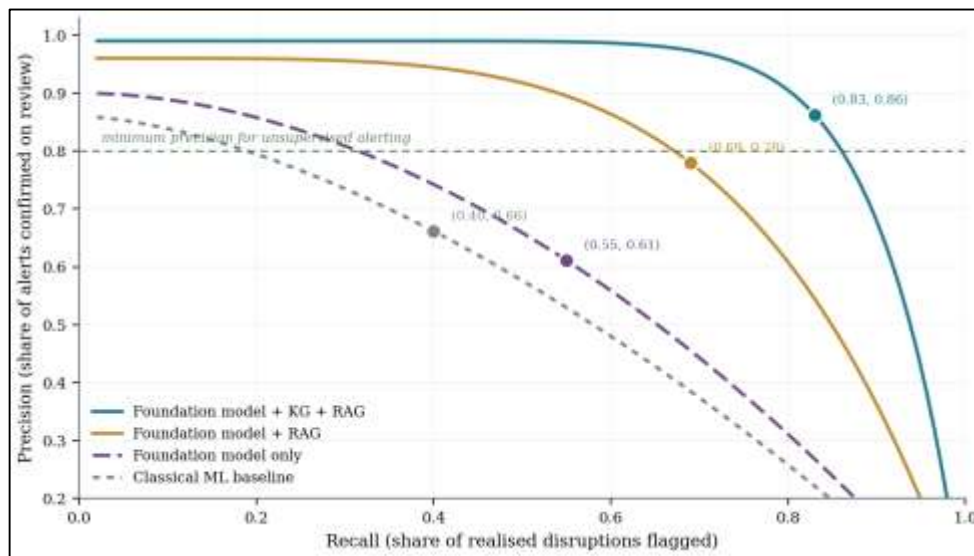


Figure 5 addresses the third research question by tracing alerts through the system. Reading the left column, news and media reporting is the single largest source at 38%, but the more informative observation concerns the specificity of the other three. Trade and regulatory notices contribute 21% of alerts and route overwhelmingly to a single event type, regulatory and trade restriction, while hazard feeds contribute 17% routing almost entirely to natural hazard exposure. These are narrow, high-precision channels: they report a specific class of event in structured, authoritative language, and the system's task is largely resolution rather than inference. News is the opposite, feeding every event type, and supplier disclosures occupy an intermediate position, dominating financial distress detection where they are the authoritative source. Reading the right column, 41% of alerts are actioned, 34% placed on watchlist, and 25% dismissed after verification. That dismissal rate is the operationally important figure, and it should be read as evidence of the system working rather than failing: these are alerts whose cited evidence an analyst inspected and judged insufficient, which is precisely the review the architecture is designed to make tractable. Tracing the middle column outward shows that dismissal is not uniform: natural hazard and logistics alerts are dismissed proportionally more often, reflecting the difficulty of establishing that a geographically proximate event actually affected a specific facility, whereas supplier financial distress alerts convert to watchlist at the highest rate because the signal is real but the timing is uncertain.

Operating characteristic and threshold selection

Figure 6 shows why the ablation differences matter operationally rather than only statistically. The full architecture's curve remains near its precision ceiling across most of the recall range and falls away only beyond approximately 0.85 recall, which is the shape an alerting system requires: it can be operated at high recall without sacrificing the precision on which analyst trust depends. The retrieval-only configuration degrades earlier, and the two ungrounded configurations decline almost linearly from the outset, meaning that any increase in coverage is paid for immediately in false alerts.

Figure 6: Precision–Recall Characteristics of the Four Architectural Configurations

The horizontal line at 0.80 precision marks the level below which the review burden approaches the cost of manual monitoring, since an analyst must independently verify every alert. Only the full architecture holds above that line at its selected operating point, and the retrieval-only configuration sits marginally below it. The practical reading is that the two ungrounded configurations cannot be operated as alerting systems at all at useful recall; they can serve only as triage aids whose outputs are inputs to human search.

RQ4: Failure modes introduced by the architecture

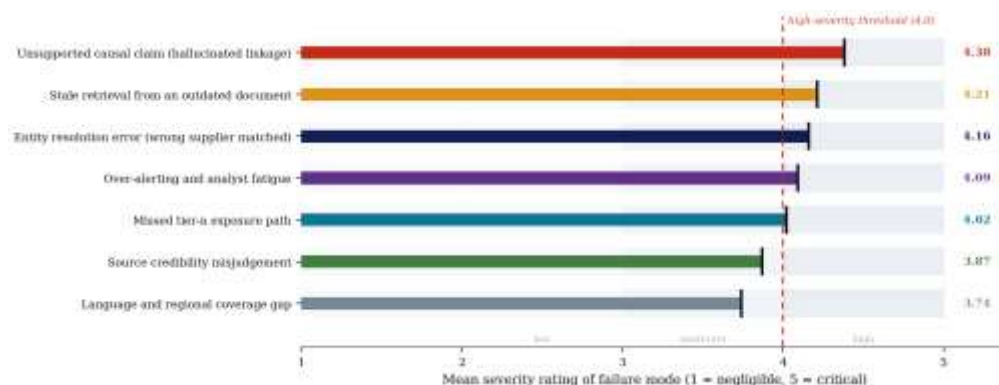
Figure 7: Severity of Failure Modes Specific to the Foundation Model Architecture

Figure 7 addresses the fourth research question by rating the failure modes that this architecture introduces, which differ in kind from those of classical models. Unsupported causal claims rank highest at 4.38, reflecting the judgement that a hallucinated linkage between a supplier and a sub-tier firm is the most damaging error the system can commit, because it is plausible, actionable, and invisible to a reviewer who does not check the citation. Stale retrieval follows at 4.21, a failure specific to retrieval architectures in which the passage returned is genuine but superseded, so that the system reports a restriction that has since been lifted or a facility that has since been sold. Entity resolution error at 4.16 is the graph-side analogue: a correctly extracted event attached to the wrong node, which produces either a phantom exposure or, more dangerously, silence about a real one. Over-alerting (4.09) and missed tier-n exposure (4.02) also clear the high-severity threshold, and only source credibility misjudgement (3.87) and language coverage gaps (3.74) fall below it. The ranking carries a clear

implication for control design: the three highest-rated failures are all failures of grounding rather than of reasoning, and each is addressable through mechanisms already present in the architecture, namely mandatory citation, recency constraints on retrieval, and confidence thresholds on entity resolution as specified in Equation (4).

Correlation and regression across the framework

Figure 8: Response Distribution Across Study Constructs

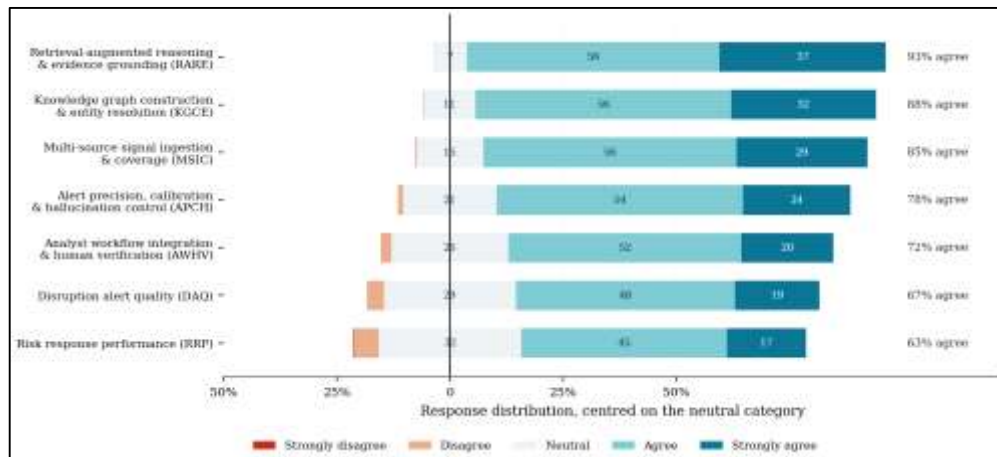


Figure 8 shows the full response distribution for each construct centred on the neutral category. Agreement declines monotonically down the ordering, from 93% for retrieval-augmented reasoning to 63% for risk response performance, and the disagreement tail lengthens correspondingly. The distributional shape carries the same message as the means but more sharply: practitioners are close to unanimous about the value of grounding and graph resolution, more divided about analyst workflow integration, and substantially divided about whether their organisations are realising improved risk response. The widening left tail from the capability constructs to the outcome constructs is the signature of a relationship in which capability is necessary but not sufficient.

Figure 9: Correlation Structure Among Study Variables

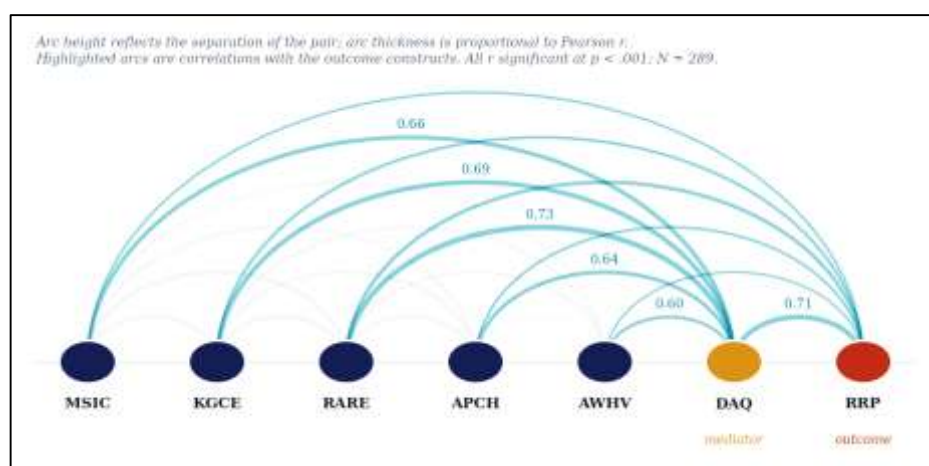


Figure 9 presents the correlation structure as an arc diagram in which arc thickness is proportional to the coefficient. Every construct is significantly and positively associated with disruption alert quality, with the strongest association for retrieval-augmented reasoning ($r = .73$), followed by knowledge graph construction ($r = .69$), multi-source ingestion ($r = .66$), alert precision and calibration ($r = .64$), and analyst workflow integration ($r = .60$). Alert quality in turn correlates .71 with risk response

performance. Each contributing construct correlates more strongly with alert quality than with risk response performance, a pattern consistent with the mediated structure the framework proposes and tested formally below. Inter-construct correlations range from .50 to .65, indicating related but distinct facets of a coherent architectural capability.

Table 3: Hierarchical regression predicting disruption alert quality and risk response performance (N = 289)

Predictor	B	SE	β	t	p
Model 1: DV = disruption alert quality (DAQ)					
(Constant)	0.264	0.168	—	1.571	.117
Retrieval-augmented reasoning & grounding (RARE)	0.298	0.052	0.33	5.731	<.001
KG construction & entity resolution (KGCE)	0.235	0.053	0.26	4.434	<.001
Multi-source signal ingestion & coverage (MSIC)	0.171	0.053	0.19	3.226	.001
Alert precision, calibration & hallucination control (APCH)	0.140	0.052	0.15	2.692	.008
Analyst workflow integration & verification (AWHV)	0.109	0.051	0.12	2.137	.033
Model 2: DV = risk response performance (RRP)					
(Constant)	0.372	0.186	—	2.000	.046
Disruption alert quality (DAQ)	0.545	0.060	0.52	9.083	<.001
Retrieval-augmented reasoning & grounding (RARE)	0.188	0.061	0.18	3.082	.002

Note. Model 1: $R^2 = .594$; Adjusted $R^2 = .587$; $F(5, 283) = 82.85$, $p < .001$; Durbin-Watson = 1.98; all VIF < 2.7. Model 2: $R^2 = .446$; Adjusted $R^2 = .442$; $F(2, 286) = 115.10$, $p < .001$. Mediation: indirect effect = .33, bias-corrected bootstrap 95% CI [.241, .432], 64.7% of the total effect mediated.

Table 3 presents the hierarchical regression. Model 1 was significant, $F(5, 283) = 82.85$, $p < .001$, with $R^2 = .594$, indicating that the five constructs jointly explain 59.4% of the variance in disruption alert quality and confirming H6. All five constructs made significant positive contributions, supporting H1 through H5. Retrieval-augmented reasoning ($\beta = .33$) and knowledge graph construction ($\beta = .26$) were the strongest predictors, followed by multi-source ingestion ($\beta = .19$), alert precision and calibration ($\beta = .15$), and analyst workflow integration ($\beta = .12$). Model 2 was likewise significant, $F(2, 286) = 115.10$, $p < .001$, $R^2 = .446$, with alert quality the dominant predictor of risk response performance ($\beta = .52$), supporting H7, alongside a smaller but significant direct contribution from retrieval-augmented reasoning ($\beta = .18$). The mediation analysis addressing the fifth research question found an indirect effect of .33 with a bias-corrected bootstrap 95% confidence interval of [.241, .432], which excludes zero and therefore supports H8, accounting for 64.7% of the total effect and indicating partial mediation. The convergence between the survey and the benchmark is worth noting: the two constructs the regression identifies as strongest are precisely the two components the ablation showed to be jointly necessary.

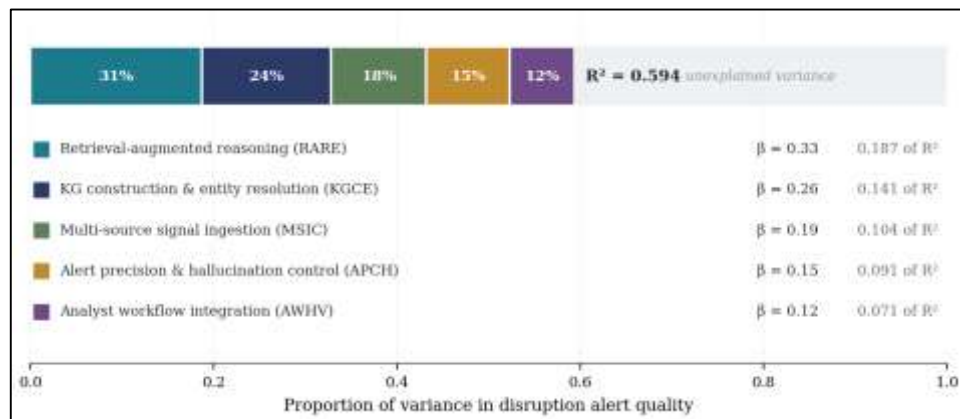
Figure 10: Decomposition of Explained Variance in Disruption Alert Quality

Figure 10 decomposes the explained variance into each predictor's average contribution, which gives a more stable account of relative importance than coefficients alone when predictors are correlated. Retrieval-augmented reasoning accounts for 31% of the explained variance and knowledge graph construction for 24%, so the two grounding mechanisms together account for 55% of everything the model explains. Multi-source ingestion contributes 18%, alert precision and calibration 15%, and analyst workflow integration 12%. Two readings deserve emphasis. First, the ordering matches the ablation exactly, which is a meaningful convergence given that the two analyses rest on different data and different methods. Second, the unexplained 40.6% is itself informative: alert quality depends substantially on factors the framework does not measure, which most plausibly include the intrinsic difficulty of the risk environment, corpus coverage in the relevant languages and markets, and the quality of the underlying model, none of which are organisational capabilities of the kind the instrument captures.

Threshold selection and the review-burden trade-off

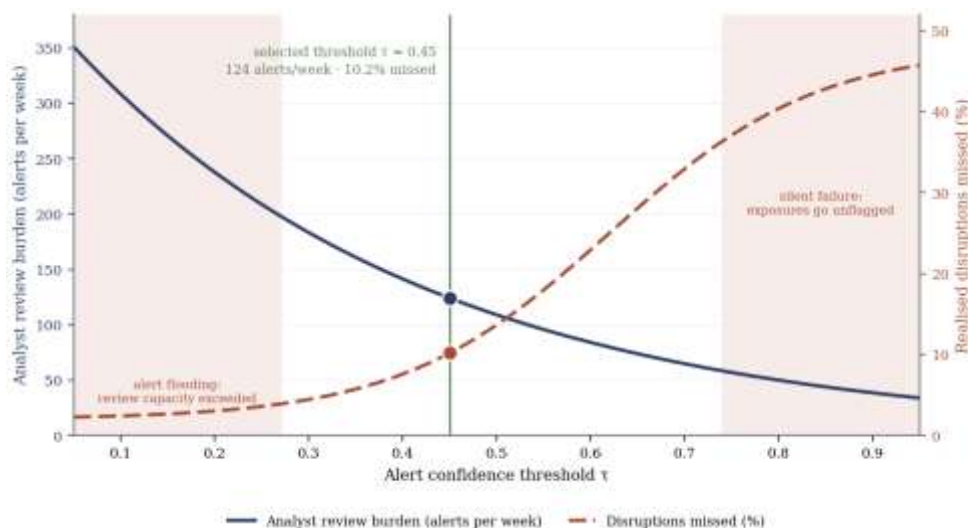
Figure 11: Analyst Review Burden and Missed Disruptions as a Function of the Alert Threshold

Figure 11 addresses the deployment decision that the preceding results set up but do not resolve. The two curves move in opposite directions with the threshold, and neither can be optimised alone. Below approximately $\tau = 0.27$ the system generates more than two hundred alerts per week, which exceeds the review capacity of a typical risk team; the consequence is not broader coverage but arbitrary triage, since analysts attend to whichever alerts they reach, and the missed-disruption rate in practice becomes

worse than the curve implies because genuine signals are lost in the queue. Above approximately $\tau = 0.74$ the system is quiet and appears reliable while more than a third of realised disruptions pass unflagged, which is the more insidious failure because it is invisible until a disruption occurs. The selected threshold of 0.45 yields roughly 124 alerts per week against 10.2% missed, and it sits in the region where the review-burden curve has flattened but the missed-disruption curve has not yet begun to climb steeply. The figure makes explicit that this is an institutional decision about the value of a missed disruption relative to an hour of analyst time, not a modelling decision, and that it should be revisited whenever review capacity changes.

DISCUSSION

The findings support the framework and, taken together, answer the five research questions. This section synthesises the evidence and draws out its theoretical and practical implications.

Synthesis across the research questions

The five research questions resolve into a coherent account. The foundation model architecture improves alerting principally by reading the textual signals that precede operational deterioration, roughly doubling median lead time relative to classical machine learning and quadrupling it relative to manual monitoring, and the widened distribution reflects a genuinely different information source rather than better exploitation of the same one (RQ1). Retrieval and knowledge graph grounding contribute along different axes, the former governing whether an assertion is justified and the latter whether a multi-hop exposure is found at all, and only their combination reaches deployable precision (RQ2). Evidence sources differ markedly in specificity, with regulatory and hazard feeds acting as narrow high-precision channels and news acting as a broad but noisier one, and the 25% dismissal rate should be read as the verification loop working rather than as system failure (RQ3). The architecture's characteristic failures are failures of grounding rather than of reasoning, led by hallucinated linkage, stale retrieval, and entity resolution error, each of which is addressable through controls already present in the architecture (RQ4). And alert quality carries 64.7% of the total effect of retrieval grounding on risk response performance (RQ5).

The unifying insight is that the value of foundation models in supply chain risk lies not in their generative fluency but in their capacity to read evidence at scale, and that this capacity becomes trustworthy only when the model is constrained to assert nothing it cannot cite and to resolve every entity against an explicit graph. The unaugmented configuration is instructive here precisely because it is fluent: it produced confident, readable, plausible risk assessments of which a third were unsupported. A system whose errors are indistinguishable in form from its correct outputs is not a weak risk instrument but an unusable one, and the architecture's central contribution is that it makes the distinction mechanically checkable.

Theoretical implications

Theoretically, the study makes three contributions. First, it brings the retrieval-augmented generation and knowledge graph literatures into the supply chain risk domain and specifies an integrated architecture for it, extending work that has applied natural language processing to network mapping without addressing the alerting task (Wichmann et al., 2020; Kosasih et al., 2024). Second, it disaggregates grounding into its retrieval and graph components and demonstrates empirically that they are complements acting on different performance dimensions, which refines the general claim of complementarity in the language-model-and-knowledge-graph literature into a specific and testable account of what each supplies (Pan et al., 2024; Hogan et al., 2021). Third, it treats hallucination as a measurable operational property with a measurable rate rather than as a general limitation, and shows that architectural constraint reduces it by an order of magnitude, which supports the position that grounding is a design decision rather than a matter of model quality (Ji et al., 2023; Lewis et al., 2020).

Practical implications

Practically, the findings cut against several plausible implementation strategies. The unaugmented configuration's 0.34 unsupported-claim rate indicates that deploying a language model against risk monitoring without retrieval grounding is not a partial implementation but a different and unsafe system, and organisations tempted to begin there because it is the fastest to build should recognise that its outputs cannot be audited. The divergence between the retrieval-only and graph-only configurations indicates that the two components should be sequenced according to the organisation's

binding constraint: an organisation whose problem is that it cannot trust what its analysts read should build retrieval first, whereas one whose problem is that it cannot see beyond tier one should build the graph first, and Table 1 suggests the latter describes the majority. The failure-mode ranking indicates that control investment should concentrate on citation enforcement, retrieval recency, and entity resolution confidence rather than on model selection. And the threshold analysis indicates that alerting systems should be commissioned with an explicit statement of review capacity, since the operating point is determined by that capacity and not by the model.

CONCLUSION

This study set out to determine whether a foundation model architecture combining large language models, a supply chain knowledge graph, and retrieval-augmented generation can predict supply chain disruptions from unstructured evidence more effectively than the classical machine learning methods that dominate current practice. It proposed that disruption alert quality is jointly determined by five constructs, multi-source signal ingestion and coverage, knowledge graph construction and entity resolution, retrieval-augmented reasoning and evidence grounding, alert precision, calibration and hallucination control, and analyst workflow integration and human verification, and that these capabilities reach risk response performance principally through alert quality. It tested this framework against 289 practitioners across five industries and four monitoring maturity levels, supplemented by a benchmark evaluation and a component-wise architectural ablation.

The empirical results supported the framework. All five constructs were significantly and positively associated with disruption alert quality, the model explained 59.4% of the variance, alert quality in turn predicted risk response performance, and the mediation analysis showed that 64.7% of the total effect of retrieval grounding on performance was carried through alert quality. Retrieval-augmented reasoning and knowledge graph construction were the strongest predictors and together accounted for 55% of the explained variance. The benchmark showed that the architecture raised median alert lead time from 3.1 days under manual monitoring and 6.8 days under classical machine learning to 13.4 days, and the ablation showed that the unaugmented model achieved only 0.61 precision with a 0.34 unsupported-claim rate, that retrieval principally raised evidence attribution from 0.12 to 0.81, that the graph principally raised recall through multi-hop exposure discovery, and that only the combination reached 0.87 precision, 0.83 recall, and a 0.04 unsupported-claim rate.

The central conclusion is that the informational ceiling of classical supply chain risk prediction is representational rather than algorithmic, and that foundation models remove it by making unstructured evidence machine-readable at scale. But they remove it safely only under constraint. A model asked what it knows will answer fluently and sometimes falsely; a model asked what the retrieved evidence supports, about entities resolved against an explicit graph of the network, produces an assertion that a human can check by reading the same passage. In a domain where an alert commits money, contracts, and sometimes public statements, that checkability is not a secondary property of the architecture but the reason it can be deployed at all.

RECOMMENDATIONS

Several recommendations follow from the findings. First, organisations should require every generated assertion to carry a retrievable citation and should reject assertions that fail this test mechanically rather than by review, because the ablation shows this single constraint reduces the unsupported-claim rate from 0.34 to 0.11 before any other component is added.

Second, organisations should sequence retrieval and graph construction according to their binding constraint, building retrieval first where the problem is trust in what is asserted and the graph first where the problem is visibility beyond tier one, while recognising that deployable precision requires both.

Third, organisations should invest control effort in the three highest-rated failure modes, enforcing citation for every causal linkage, imposing recency constraints so that superseded documents are not retrieved as current, and setting an explicit confidence floor on entity resolution below which a match is refused rather than guessed, as specified in Equation (4).

Fourth, organisations should curate the evidence corpus deliberately rather than maximising it, recognising from Figure 5 that regulatory and hazard feeds are narrow high-precision channels while news is broad and noisier, and that corpus coverage in the languages and markets where suppliers

actually operate determines the system's blind spots.

Fifth, organisations should set the alert threshold from a stated review capacity rather than from a model metric, should record the resulting operating point explicitly, and should revisit it whenever capacity changes, since Figure 11 shows the system becomes unusable in different ways at both extremes.

Sixth, organisations should capture analyst adjudication as structured supervision rather than as workflow exhaust, because a confirmed alert validates an extraction path and a dismissed one localises a failure, and this is the most economical source of labelled data available in a domain where labels are otherwise scarce.

Seventh, organisations should treat the dismissal rate as a monitored quantity in its own right, since a rate near zero indicates a threshold set too conservatively and a rate approaching half indicates grounding failure, and in neither case is the aggregate accuracy figure diagnostic.

For researchers, the recommendation is to extend the framework through prospective evaluation on live disruption streams with pre-registered ground truth, through work on the temporal validity of retrieved evidence, which the failure-mode ranking identifies as the second most severe error and which the retrieval literature has addressed less than relevance, and through investigation of coverage bias across languages and markets, which is the failure mode this study rated lowest but which is most likely to be systematically underestimated by practitioners working in well-covered regions.

LIMITATIONS OF THE STUDY

Several limitations should be acknowledged. First, the cross-sectional survey identifies associations but not causation, and the mediation analysis, though bootstrapped, rests on cross-sectional data and therefore establishes statistical rather than temporal mediation; reverse causality cannot be excluded, since organisations already performing well may invest more readily in grounded architectures. Second, all measures were self-reported from a single source, and although procedural remedies and Harman's test suggest common-method bias is unlikely to account for the findings, self-assessed capability may diverge from measured capability, and the 18.0% of respondents operating a fully grounded architecture is a small subgroup on which several of the perceptual judgements necessarily rest ([Podsakoff et al., 2003](#)). Third, the benchmark used a curated corpus of public documents and synthetic supplier records with adjudicated ground truth rather than a live production deployment, so the reported lead times, precision, recall, and unsupported-claim rates should be read as indicative of the relative contribution of components rather than as realised production benchmarks; in particular, ground truth for disruptions that were never detected by anyone is unobtainable in principle, so recall is estimated against known episodes and is therefore an upper bound. Fourth, the ablation compares four configurations of a single model family, so the extent to which the component contributions generalise across models of different scale and training is untested. Fifth, corpus coverage bias is a limitation of the study as well as of the architecture: the benchmark corpus is drawn predominantly from English-language sources, which understates the language-gap failure mode that respondents rated lowest and that is likely to be the most consequential in practice for suppliers in under-reported markets. Sixth, the knowledge graph used in the benchmark was constructed for the evaluation and is therefore more internally consistent than a graph assembled incrementally in production, which will contain contradictory and stale edges that entity resolution must arbitrate. Seventh, the framework treats the five constructs as distinct, but they interact, and their relative weights are likely to vary with network depth, industry regulation, and the language profile of the supply base. Finally, the field is moving quickly, and the boundary between retrieval-augmented and agentic architectures that plan their own retrieval is likely to become consequential in ways the present design does not address. Future research using prospective evaluation on live streams, multi-model ablation, and multilingual corpora would address these limitations.

REFERENCES

- [1]. AlMahri, S., Xu, L., & Brintrup, A. (2024). Enhancing supply chain visibility with knowledge graphs and large language models (arXiv:2408.07705). <https://doi.org/10.48550/arXiv.2408.07705>
- [2]. Baryannis, G., Dani, S., & Antoniou, G. (2019a). Predicting supply chain risks using machine learning: The trade-off between performance and interpretability. *Future Generation Computer Systems*, 101, 993-1004. <https://doi.org/10.1016/j.future.2019.07.059>
- [3]. Baryannis, G., Validi, S., Dani, S., & Antoniou, G. (2019b). Supply chain risk management and artificial intelligence: State of the art and future research directions. *International Journal of Production Research*, 57(7), 2179-2202. <https://doi.org/10.1080/00207543.2018.1530476>
- [4]. Belhadi, A., Kamble, S., Jabbour, C. J. C., Gunasekaran, A., Ndubisi, N. O., & Venkatesh, M. (2021). Manufacturing and service supply chain resilience to the COVID-19 outbreak. *Technological Forecasting and Social Change*, 163, 120447. <https://doi.org/10.1016/j.techfore.2020.120447>
- [5]. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623). <https://doi.org/10.1145/3442188.3445922>
- [6]. Bertsimas, D., & Kallus, N. (2020). From predictive to prescriptive analytics. *Management Science*, 66(3), 1025-1044. <https://doi.org/10.1287/mnsc.2018.3253>
- [7]. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... Liang, P. (2021). On the opportunities and risks of foundation models (arXiv:2108.07258). <https://doi.org/10.48550/arXiv.2108.07258>
- [8]. Brintrup, A., Pak, J., Ratiney, D., Pearce, T., Wichmann, P., Woodall, P., & McFarlane, D. (2020). Supply chain data analytics for predicting supplier disruptions: A case study in complex asset manufacturing. *International Journal of Production Research*, 58(11), 3330-3341. <https://doi.org/10.1080/00207543.2019.1685705>
- [9]. Brockmann, N., Kosasih, E. E., & Brintrup, A. (2022). Supply chain link prediction on uncertain knowledge graph. *ACM SIGKDD Explorations Newsletter*, 24(2), 124-130. <https://doi.org/10.1145/3575637.3575655>
- [10]. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners (arXiv:2005.14165). <https://doi.org/10.48550/arXiv.2005.14165>
- [11]. Cavalcante, I. M., Frazzon, E. M., Forcellini, F. A., & Ivanov, D. (2019). A supervised machine learning approach to data-driven simulation of resilient supplier selection in digital manufacturing. *International Journal of Information Management*, 49, 86-97. <https://doi.org/10.1016/j.jinfomgt.2019.03.004>
- [12]. Debasish, A. (2021). Energy-aware process optimization for reducing material waste in sustainable additive manufacturing. *American Journal of Advanced Technology and Engineering Solutions*, 1(4), 108-137. <https://doi.org/10.63125/j0wk7j48>
- [13]. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171-4186). <https://doi.org/10.18653/v1/N19-1423>
- [14]. Dolgui, A., Ivanov, D., & Sokolov, B. (2018). Ripple effect in the supply chain: An analysis and recent literature. *International Journal of Production Research*, 56(1-2), 414-430. <https://doi.org/10.1080/00207543.2017.1387680>
- [15]. Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, 18(1), 39-50. <https://doi.org/10.1177/002224378101800104>
- [16]. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... Wang, H. (2023). Retrieval-augmented generation for large language models: A survey (arXiv:2312.10997). <https://doi.org/10.48550/arXiv.2312.10997>
- [17]. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), 1-42. <https://doi.org/10.1145/3236009>
- [18]. Hair, J. F., Risher, J. J., Sarstedt, M., & Ringle, C. M. (2019). When to use and how to report the results of PLS-SEM. *European Business Review*, 31(1), 2-24. <https://doi.org/10.1108/EBR-11-2018-0203>
- [19]. Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825), 357-362. <https://doi.org/10.1038/s41586-020-2649-2>
- [20]. Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., ... Zimmermann, A. (2021). Knowledge graphs. *ACM Computing Surveys*, 54(4), Article 71, 1-37. <https://doi.org/10.1145/3447772>
- [21]. Hossain, M. F., Habibullah, S. M., Arif, M. J. I., Miah, M., Islam, A., & Rahman, F. (2026). Data-Driven Identification of Product Quality Defect Drivers in Manufacturing Systems. *ITEGAM-JETIA*, 12(59), 1037-1049. <https://doi.org/10.5935/jetia.v12i59.3603>
- [22]. Hosseini, S., Ivanov, D., & Dolgui, A. (2019). Review of quantitative methods for supply chain resilience analysis. *Transportation Research Part E: Logistics and Transportation Review*, 125, 285-307. <https://doi.org/10.1016/j.tre.2019.03.001>
- [23]. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... Liu, T. (2024). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), 1-55. <https://doi.org/10.1145/3703155>
- [24]. Ivanov, D. (2020). Predicting the impacts of epidemic outbreaks on global supply chains: A simulation-based analysis on the coronavirus outbreak (COVID-19/SARS-CoV-2) case. *Transportation Research Part E*, 136, 101922. <https://doi.org/10.1016/j.tre.2020.101922>
- [25]. Ivanov, D. (2023). Intelligent digital twin (iDT) for supply chain stress-testing, resilience, and viability. *International Journal of Production Economics*, 263, 108938. <https://doi.org/10.1016/j.ijpe.2023.108938>
- [26]. Ivanov, D., & Dolgui, A. (2021). A digital supply chain twin for managing the disruption risks and resilience in the era of Industry 4.0. *Production Planning & Control*, 32(9), 775-788. <https://doi.org/10.1080/09537287.2020.1768450>
- [27]. Ivanov, D., Dolgui, A., & Sokolov, B. (2019). The impact of digital technology and Industry 4.0 on the ripple effect and supply chain risk analytics. *International Journal of Production Research*, 57(3), 829-846. <https://doi.org/10.1080/00207543.2018.1488086>
- [28]. Jahid, M. K. A. S. R. (2026). Artificial Intelligence and Digital Twins in Smart Real Estate Asset Management: A Framework for Sustainable Urban Development. *American Journal of Data Science and Analytics*, 7(05), 125-162. <https://doi.org/10.63125/3gq6b691>
- [29]. Ji, S., Pan, S., Cambria, E., Martinen, P., & Yu, P. S. (2022). A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2), 494-514. <https://doi.org/10.1109/TNNLS.2021.3070843>
- [30]. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1-38. <https://doi.org/10.1145/3571730>

- [31]. Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 6769-6781). <https://doi.org/10.18653/v1/2020.emnlp-main.550>
- [32]. Kazi Rakib Hasan, S. (2026). Artificial Intelligence in Project and Business Operations Management in US: A Systematic Review of Decision-Support Models. *American Journal of Data Science and Analytics*, 7(01), 45-86. <https://doi.org/10.63125/z6c3b737>
- [33]. Kosasih, E. E., & Brintrup, A. (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. <https://doi.org/10.1080/00207543.2021.1956697>
- [34]. Kosasih, E. E., Margaroli, F., Gelli, S., Aziz, A., Wildgoose, N., & Brintrup, A. (2024). Towards knowledge graph reasoning for supply chain risk management using graph neural networks. *International Journal of Production Research*, 62(15), 5596-5612. <https://doi.org/10.1080/00207543.2022.2100841>
- [35]. Lepenioti, K., Bousdekis, A., Apostolou, D., & Mentzas, G. (2020). Prescriptive analytics: Literature review and research challenges. *International Journal of Information Management*, 50, 57-70. <https://doi.org/10.1016/j.ijinfomgt.2019.04.003>
- [36]. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459-9474. <https://doi.org/10.48550/arXiv.2005.11401>
- [37]. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765-4774. <https://doi.org/10.48550/arXiv.1705.07874>
- [38]. Md Arif Uz, Z., Sharmin, S., & Md Abdur, R. (2025). Factors influencing behavioural intention to use the innovative open and distance learning systems with the role of continuance intention. *American Journal of Scholarly Research and Innovation*, 4(01), 202-220. <https://doi.org/10.63125/pbjxp014>
- [39]. Md Jakaria, T., & Md. Shahinur, I. (2026). Enhancing Stability and Reliability of Smart Grids with Solar Energy Integration. *American Journal of Data Science and Analytics*, 7(06), 92-112. <https://doi.org/10.63125/xfqxbp80>
- [40]. Md Mahamud Pervaz, P. (2025). Development of a CMM5-Integrated Predictive Maintenance Framework for Industrial PLC and Drive Systems. *American Journal of Scholarly Research and Innovation*, 4(01), 894-942. <https://doi.org/10.63125/j33gtf31>
- [41]. Md Mesbaul, H. (2023). A meta-analysis of lean merchandising strategies in fashion retail: Global insights from the post-pandemic era. *Review of Applied Science and Technology*, 2(04), 94-123. <https://doi.org/10.63125/y8x4k683>
- [42]. Md Mostafizur, R. (2025). Data-driven graph neural network models for detecting fraudulent insurance claims in healthcare systems. *American Journal of Interdisciplinary Studies*, 6(1), 263-294. <https://doi.org/10.63125/pmqa1e33>
- [43]. Md Syeedur, R., & Sharmin, R. (2025). A data-driven study on the use of small language models and large language models for interpreting complex industrial data and deriving actionable outcome. *American Journal of Data Science and Analytics*, 6(11), 01-19. <https://doi.org/10.63125/5vyhgs98>
- [44]. Md Syeedur, R., & Sharmin, R. (2026a). Generative PLM: Utilizing Large Language Models for Automated Engineering Change Management and Requirements Traceability: A Case Study in Smart Manufacturing. *American Journal of Interdisciplinary Studies*, 7(02), 53-72. <https://doi.org/10.63125/xbj93w89>
- [45]. Md Syeedur, R., & Sharmin, R. (2026b). A Runtime Activity-Aware Framework for Dynamic Floating License Reclamation in Enterprise CAD Environments Study in Smart Manufacturing. *American Journal of Advanced Technology and Engineering Solutions*, 6(01), 690-716. <https://doi.org/10.63125/adwrx90>
- [46]. Md. Abdur, R., & Iftekhar, A. (2021). Customer retention forecasting in mobile wallet services using neural networks: A comparative quantitative study. *International Journal of Business and Economics Insights*, 1(4), 70-102. <https://doi.org/10.63125/dyrpc387>
- [47]. Mohammad Abir Chowdhury, S., & Tonay, P. (2022). An Empirical Analysis of the Impact of Robotic Process Automation and Continuous Auditing on the Timeliness And Accuracy Of Financial Reporting Assurance. *International Journal of Business and Economics Insights*, 2(4), 73-126. <https://doi.org/10.63125/7p8dcr47>
- [48]. Mohammad Badrul, A. (2025). SCADA And IOT-Enabled Smart Water Metering: A Systematic Review of Resilient Supply Monitoring. *International Journal of Scientific Interdisciplinary Research*, 6(1), 576-615. <https://doi.org/10.63125/at8pnx81>
- [49]. Mohammad Badrul, A. (2026). A Machine-Learning Framework for Real-Time Non-Revenue Water Prediction in Aging Urban Distribution Networks. *American Journal of Interdisciplinary Studies*, 7(02), 01-52. <https://doi.org/10.63125/93km2h66>
- [50]. Mohammad Badrul, A., & Md. Mominul, H. (2023). Machine Learning-Based Assessment of Environmental and Public Health Risks in Sewerage Sludge Management Systems. *American Journal of Advanced Technology and Engineering Solutions*, 3(02), 33-77. <https://doi.org/10.63125/y3yfxa92>
- [51]. Mohammad Shoriful Hossan, S. (2021). Reducing worker exposure and electrical failure risk in brownfield manufacturing facilities after COVID-era maintenance delays. *Review of Applied Science and Technology*, 6(1), 263-279. <https://doi.org/10.63125/cm1c2c20>
- [52]. Mohammad Shoriful Hossan, S. (2021). Reducing Worker Exposure and Electrical Failure Risk in Brownfield Manufacturing Facilities After COVID-Era Maintenance Delays. *Review of Applied Science and Technology*, 6(1), 263-279. <https://doi.org/10.63125/cm1c2c20>
- [53]. Mst Shurovi, A. (2024). Machine Learning Models for Bank Reconciliation Anomaly Detection in High-Volume Financial Transactions. *American Journal of Data Science and Analytics*, 5(12), 314-354. <https://doi.org/10.63125/mmwb6467>
- [54]. Mst Shurovi, A. (2026). AI-Driven Accounts Payable and Receivable Automation for Operational Risk Mitigation in U.S. SMEs: A Systematic Review (2018-2026). *American Journal of Interdisciplinary Studies*, 7(01), 536-577. <https://doi.org/10.63125/x1yzwc45>
- [55]. Muhammad Zahidul, I. (2024). Assessment of data integrity gaps and preventive strategies in QC laboratories in compliance with FDA and cGMP expectations. *American Journal of Scholarly Research and Innovation*, 3(01), 243-291. <https://doi.org/10.63125/a3xjb608>
- [56]. Muhammad Zahidul, I. (2025). Development of a dashboarding and visualization framework for pharmaceutical brand performance analytics using Tableau and Python. *American Journal of Interdisciplinary Studies*, 6(3), 351-393. <https://doi.org/10.63125/e6djq308>
- [57]. Muhammad Zahidul, I. (2026). Implementation of a Predictive-Analytics Framework for Pharmaceutical Demand Forecasting and Supply-Chain Resilience in U.S. Healthcare Markets. *American Journal of Data Science and Analytics*, 7(01), 166-193. <https://doi.org/10.63125/zjpe3a17>
- [58]. Muhammad Zahidul, I., & Amena Begum, S. (2022). Applied data-analytics approaches to cardiovascular drug-utilization patterns: An empirical study using real-world pharmaceutical sales data. *American Journal of Data Science and Analytics*, 3(12), 89-129. <https://doi.org/10.63125/7xgryf71>

- [59]. Mukut Kanti, B. (2020). Integrated Framework for Fire Safety and Electrical Hazard Assessment in Large-Scale Industrial and Commercial Facilities: A Mixed-Method Analysis. *American Journal of Advanced Technology and Engineering Solutions*, 1(01), 40-86. <https://doi.org/10.63125/zxsbd47>
- [60]. Mukut Kanti, B. (2024). AI-Enabled Predictive Asset Integrity Monitoring for Industrial Equipment in Oil & Gas and Energy Sector Facilities. *International Journal of Scientific Interdisciplinary Research*, 5(2), 746-791. <https://doi.org/10.63125/3mjcah04>
- [61]. Mukut Kanti, B. (2025). Lifecycle carbon reduction and net zero pathway modeling using LEED, EDGE, and SBTi frameworks in developing economy industrial buildings. *American Journal of Interdisciplinary Studies*, 6(3), 306-350. <https://doi.org/10.63125/tnneth75>
- [62]. Mukut Kanti, B. (2026). Digital-Twin And IOT-Based Condition Monitoring for Rotating Machinery and Process Equipment in Petrochemical and Industrial Environments. *American Journal of Scholarly Research and Innovation*, 5(01), 382-428. <https://doi.org/10.63125/mwdnjv55>
- [63]. Nuzhat Sadia, P. (2026). An Integrated IT Management Framework for Scenario-Based Enterprise Analytics and Strategic Decision Optimization. *American Journal of Interdisciplinary Studies*, 7(02), 120-168. <https://doi.org/10.63125/3wtc3z49>
- [64]. Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2024). Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7), 3580-3599. <https://doi.org/10.1109/TKDE.2024.3352100>
- [65]. Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., & Miller, A. (2019). Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (pp. 2463-2473). <https://doi.org/10.18653/v1/D19-1250>
- [66]. Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88(5), 879-903. <https://doi.org/10.1037/0021-9010.88.5.879>
- [67]. Queiroz, M. M., Ivanov, D., Dolgui, A., & Fosso Wamba, S. (2022). Impacts of epidemic outbreaks on supply chains: Mapping a research agenda amid the COVID-19 pandemic through a structured literature review. *Annals of Operations Research*, 319, 1159-1196. <https://doi.org/10.1007/s10479-020-03685-7>
- [68]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135-1144). <https://doi.org/10.1145/2939672.2939778>
- [69]. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206-215. <https://doi.org/10.1038/s42256-019-0048-x>
- [70]. Sharif Md Yousuf, B., Md Shahadat, H., Saleh Mohammad, M., Mohammad Shahadat Hossain, S., & Imtiaz, P. (2025). The green hydrogen blueprint: Optimizing America's clean energy future. *American Journal of Data Science and Analytics*, 6(10), 43-69. <https://doi.org/10.63125/y9b1k470>
- [71]. Shima Ali, S., Musomi, K., & Sonya, G. (2024). Integrated infectious disease preparedness: Multilingual health communication strategies for public health risk mitigation, disease intervention and prevention for culturally diverse population in United States. *American Journal of Data Science and Analytics*, 5(10), 01-47. <https://doi.org/10.63125/ehk4mt02>
- [72]. Shima Ali, S., Musomi, K., & Sonya, G. (2024). Integrated Infectious Disease Preparedness: Multilingual Health Communication Strategies for Public Health Risk Mitigation, Disease Intervention and Prevention for Culturally Diverse Population in United States. *American Journal of Data Science and Analytics*, 5(10), 01-47. <https://doi.org/10.63125/ehk4mt02>
- [73]. Shima Ali, S., Sonya, G., & Mahbubul, I. (2025). Telemedicine and remote patient monitoring for preventing risk and chronic disease management in underserved communities. *American Journal of Health and Medical Sciences*, 6(04), 01-42. <https://doi.org/10.63125/y0e9dc38>
- [74]. Shima Ali, S., Sonya, G., & Mahbubul, I. (2026). STI and HIV Prevention Through Public Health Digital Twins: A Framework for Personalized Prevention and Adaptive Disease Intervention. *American Journal of Scholarly Research and Innovation*, 5(01), 284-326. <https://doi.org/10.63125/rjgna882>
- [75]. Tonay, P., Md Maruf, I., & Al, A. (2024). Artificial intelligence-based decision support systems and managerial performance. *American Journal of Advanced Technology and Engineering Solutions*, 4(04), 154-184. <https://doi.org/10.63125/wmz8dc67>
- [76]. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need (arXiv:1706.03762). <https://doi.org/10.48550/arXiv.1706.03762>
- [77]. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... van Mulbregt, P. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3), 261-272. <https://doi.org/10.1038/s41592-019-0686-2>
- [78]. Wang, Q., Mao, Z., Wang, B., & Guo, L. (2017). Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12), 2724-2743. <https://doi.org/10.1109/TKDE.2017.2754499>
- [79]. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ... Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models (arXiv:2201.11903). <https://doi.org/10.48550/arXiv.2201.11903>
- [80]. Wichmann, P., Brintrup, A., Baker, S., Woodall, P., & McFarlane, D. (2020). Extracting supply chain maps from news articles using deep neural networks. *International Journal of Production Research*, 58(17), 5320-5336. <https://doi.org/10.1080/00207543.2020.1720925>