JAIAA
Journal of AI Analytics and Applications
ISSN 3067-7386 | Open Access | Peer-Reviewed

RESEARCH ARTICLE
Vol. 3, No. 3 (2025), pp. 75–94
DOI: 10.63646/jaiaa.2025.030305

Constraint-Guided LLM Reasoning for Causal Risk Propagation in Cyber-Physical Industrial Systems

Marta Fabbri[1], Giorgio Bianchi[2], Elena Grecoo[3,*]

[1] Department of Information Engineering, University of Brescia, Brescia, Italy
[2] Department of Industrial Engineering, University of Salerno, Fisciano, Italy
[3] Department of Computer Science, University of Bari Aldo Moro, Bari, Italy
* Corresponding author: elena.greco@uniba.it

ARTICLE INFO
Received
April 08, 2025
Revised
June 11, 2025
Accepted
August 22, 2025
Available Online
September 30, 2025
DOI
10.63646/jaiaa.2025.030305

Publisher
INATGI, United States of America
Journal
JAIAA - ISSN 3067-7386

Abstract
Cyber-physical industrial systems integrate physical equipment, embedded control, operational software, business workflows, and human decision making. When cyberattacks disturb this integration, risk rarely remains inside a single device or a single control loop. It moves across process variables, unsafe control actions, task failures, alarms, operator responses, and production objectives. Large language models offer a promising way to accelerate causal risk analysis because they can convert heterogeneous engineering descriptions into structured candidate scenarios. However, unconstrained generation may also create unsupported nodes, illegal causal jumps, ambiguous terminology, and outputs that cannot be mapped into quantitative risk models. This paper develops a constraint-guided LLM reasoning framework for causal risk propagation in cyber-physical industrial systems. The framework combines five constraint families: structural consistency, task-topology validity, semantic typing, evidence traceability, and computable output formatting. These constraints guide the model from system knowledge and attack evidence toward auditable causal chains and Bayesian-network-ready tables. A synthetic benchmark involving 24 cyber-physical risk scenarios across four industrial settings is used to evaluate the framework. The results show that constraint-guided reasoning improves causal validity from 0.58 to 0.82 compared with unconstrained prompting, increases Bayesian-network capability from 0.43 to 0.76, and reduces hallucinated or unsupported causal items by more than half. Adding retrieval support and review calibration further improves causal validity to 0.92. The study contributes an actionable design for integrating LLM reasoning into industrial safety analytics without treating generated text as authoritative evidence. It also provides implementation guidance for AI-assisted STPA, FMEA, process-risk modeling, and risk-informed decision support in industrial cyber-physical environments.

Keywords: Cyber-physical industrial systems; cyberattack; causal risk propagation; large language model; constraint-guided reasoning; risk analysis; Bayesian network; industrial safety analytics.

I. INTRODUCTION

Industrial cyber-physical systems are no longer isolated control islands. In contemporary process plants, manufacturing cells, water-treatment facilities, and energy infrastructures, sensors, controllers, actuators, supervisory systems, cloud services, maintenance platforms, and enterprise planning systems are connected through layered digital architectures. This connectivity enables flexible production and improved visibility, but it also creates new pathways through which cyber disturbances can become operational, safety, and business risks. A manipulated sensor value may trigger a wrong control decision; a delayed alarm may weaken operator intervention; a corrupted production message may disrupt downstream scheduling; and a false normal state may hide an emerging process deviation until the system reaches a

RESEARCH ARTICLE
Vol. 3, No. 3 (2025), pp. 75–94
DOI: 10.63646/jaiaa.2025.030305

hazardous operating condition. In these situations, risk does not travel in a simple linear path. It propagates through physical processes, control logic, task execution, information flows, and organizational responses. Cyber-physical manufacturing architecture research similarly emphasizes the link between physical assets, computation, and operational decision loops (Lee et al.,2015).

Classical risk-assessment methods provide valuable foundations for analyzing such systems. Failure mode and effects analysis is useful for decomposing component or task failures. System-theoretic process analysis is useful for modeling unsafe control actions and system-level hazards. Attack trees and related security models provide structures for representing adversarial paths. Bayesian networks support probabilistic reasoning under uncertainty. Yet each method tends to emphasize a particular analytical lens. Device-centric analysis can miss the way a local fault affects business objectives. Control-centric analysis may underrepresent task-level delays and workflow dependencies. Business-process analysis can describe activity failures but often lack explicit links to control constraints and physical hazards. The challenge is therefore not the absence of methods, but the difficulty of integrating heterogeneous methods into a coherent causal representation that remains auditable, computable, and useful for decision making. FMEA prioritization research shows that failure descriptions require structured treatment before they can support comparative assessment (Bowles and Peláez,1995). Evidential FMEA research reinforces the need to separate local failure modes from system-level consequences (Chin et al.,2009). FMEA review research further shows that risk evaluation becomes more reliable when failure modes and effects are consistently typed (Liu et al.,2013). Modified FMEA methods show that failure prioritization depends heavily on the structure of the scoring and grouping process (Sankar and Prabhu,2001).

Large language models have entered this methodological gap. They can read textual descriptions of equipment, operating procedures, control objectives, and incident scenarios; they can produce structured summaries; and they can propose plausible causal relationships. In early safety-engineering tasks, this capacity appears attractive because much of the required knowledge is stored in manuals, operating notes, hazard logs, incident reports, and expert narratives. LLMs may reduce the time needed to draft preliminary risk chains, identify candidate unsafe control actions, or align business tasks with control loops. They may also support human analysts by translating domain descriptions into tables that are easier to review and convert into causal graphs. AI survey research highlights that analytical usefulness depends on matching model capability to application context (Zhang and Lu,2021). Transformer-based attention provides the representational basis for much of the current language-model capability used in structured reasoning workflows (Vaswani et al.,2017). Few-shot learning results explain why LLMs can adapt quickly to domain prompts while still needing external control (Brown et al.,2020).

The same capability creates a serious problem. A general-purpose LLM is optimized to produce plausible language, not verified engineering evidence. In risk assessment, a plausible but unsupported causal link can be more harmful than an obviously wrong statement because it may enter a downstream quantitative model and influence risk priorities. A generated node may confuse a device failure with a system hazard. A generated edge may jump directly from a cyberattack to a safety loss without representing the intervening process state. A business loss may be mislabeled as a task failure effect. A list may be fluent but impossible to map into a Bayesian network because it does not separate nodes, states, parent relationships, or evidence sources. These errors are not minor wording defects. They change the meaning of the risk model. Concerns about fluent but weakly grounded language generation motivate the insistence on evidence-bearing outputs (Bender et al.,2021). Hallucination surveys show that fluent generation can still violate factual or structural constraints (Ji et al.,2023). Research on model memory explains why parametric knowledge should not be treated as a sufficient evidence source for safety-critical analysis (Mallen et al.,2023).

This paper argues that LLMs should not be used as free-form authors of industrial risk assessments. They should be used as constrained reasoning assistants embedded inside a verifiable analytical workflow. Constraint-guided reasoning means that the model is not simply asked to “identify risks.” It gives explicit restrictions on what objects may be created, what causal relationships are legally admissible, what semantic types must be preserved, what evidence must support each claim, and what output structure is required for downstream computation. The aim is not to eliminate human judgment. Rather, the aim is to make LLM-generated causal candidates easier to audit, revise, and translate into probabilistic risk

models. The design of these constraints is consistent with prompt engineering research that treats prompts as task specifications rather than stylistic instructions (Liu et al.,2023). Stepwise reasoning research supports decomposing industrial risk analysis into explicit intermediate stages (Wei et al.,2022).

The contribution of this paper is threefold. First, it proposes a constraint-guided LLM reasoning framework for causal risk propagation in cyber-physical industrial systems. The framework is designed around five constraint families: structural consistency, task-topology validity, semantic typing, evidence traceability, and computable output formatting. Second, it develops a benchmark evaluation using 24 synthetics but domain-informed risk scenarios across four industrial settings: pressure control in a process unit, water-treatment automation, robotic assembly, and energy microgrid coordination. Third, it provides a data analysis that compares unconstrained prompting, role-based prompting, constraint-guided prompting, retrieval-supported prompting, and review-calibrated prompting across causal validity, semantic precision, Bayesian network capability, redundancy, and hallucination control. The results show that constraints substantially improve the engineering usefulness of LLM output, especially when the output is intended to support quantitative risk inference. Recent work on large language models in blockchain-enabled supply chain finance also shows that language reasoning becomes more valuable when it is connected to governed transactional structures (Yang et al.,2025). Industry 4.0 blockchain research indicates that trusted information exchange is becoming part of the industrial control and business-process environment (Chen et al.,2024).

II. CONCEPTUAL BACKGROUND

Cyber-physical industrial systems combine physical dynamics with networked computation. Their risk profile differs from that of traditional information systems because the consequences of a cyber event may be physical, temporal, and irreversible. The loss of a database record can be restored from backup; the loss of containment, excessive pressure, or unsafe robot motion can cause immediate damage. Industrial AI therefore needs a risk language that can connect data-integrity events, process deviations, unsafe control actions, human responses, and business continuity. The industrial informatics literature has long emphasized that digital integration improves productivity only when it is aligned with control, interoperability, and safety requirements. Broad CPS security surveys support the view that industrial risk must be analyzed across cyber, physical, and control layers rather than within one isolated layer (Humayed et al.,2017). Industry 4.0 surveys describe cyber-physical systems as an organizing foundation for digitally connected industrial production (Lu,2017a). Industrial IoT studies identify interoperability, security, and latency as persistent barriers for connected industrial analytics (Sisinni et al.,2018).

Risk propagation in these systems is cross-domain. A single adverse event may begin in the cyber domain as compromised authentication, network delay, malicious configuration, or false data injection. It can enter the device domain as an abnormal sensor reading, actuator misbehavior, or controller state inconsistency. It may then enter the control domain through missing, delayed, excessive, or unsafe control action. It may reach the business domain through lost throughput, defective product, delayed delivery, or reduced energy efficiency. Finally, it may become a governance issue when operators and managers must decide whether to shut down a line, override a recommendation, or continue production under uncertainty. Modeling this chain requires both engineering precision and managerial relevance. Reviews of SCADA risk assessment also show that method choice strongly affects whether cyber threats are translated into operational consequences (Cherdantseva et al.,2016). IoT cybersecurity research further supports the need to trace how connected devices create new attack and dependency surfaces (Lu and Xu,2019). Attack-detection theory in CPSs shows that malicious interventions can appear as control abnormalities unless their causal signatures are modeled explicitly (Pasqualetti et al.,2013). Research on embedding blockchain into IoT security illustrates how device-level trust mechanisms can affect broader industrial information flows (Xu et al.,2021).

STPA contributes system-level safety vocabulary. It treats accidents as outcomes of inadequate control rather than merely component failure. This framing is particularly valuable for cyber-physical systems because a component may operate as designed while the overall control structure becomes unsafe under a particular context. FMEA contributes complementary tasks and failure vocabulary. It asks what may fail, why it may fail, and what effects may follow. When adapted from a

RESEARCH ARTICLE
Vol. 3, No. 3 (2025), pp. 75–94
DOI: 10.63646/jaiaa.2025.030305

device-centered view to a task-centered view, it can describe omitted tasks, incomplete execution, unstable task output, excessive response time, or premature execution. These two approaches become more powerful when they are linked: task failures may produce unsafe control actions; unsafe control actions may create hazards; hazards may produce safety losses; and task-level disruptions may also produce business losses. Safety-security co-engineering research supports the paper's separation of safety objects from security triggers while still linking them causally (Kriaa et al.,2015). Secure-control research shows that attacker capability and process response must be considered together (Teixeira et al.,2015).

Bayesian networks add an inference layer. A causal graph alone is a qualitative structure. A Bayesian network associates each node with states and conditional probabilities, allowing the analyst to update risk estimates when new evidence appears. This is valuable for cyber-physical risk because attacks, faults, and human interventions are uncertain. The same attack evidence may have different implications depending on whether a safety interlock is available, whether alarms are trusted, whether operators respond within a time window, and whether production conditions are near a boundary state. Bayesian reasoning supports sensitivity analysis and scenario comparison, both of which are needed for risk-informed decision making. Bayesian-network tool research demonstrates why node definition and edge validity must be settled before probabilistic inference is meaningful (Scutari,2010). Bayesian-network classification research provides a basis for understanding how causal dependencies can support probabilistic decision tasks (Friedman et al.,1997). Bayesian-network applications in dependability and maintenance show why probabilistic graphs are suitable for operational risk reasoning (Weber et al.,2012).

LLMs add a natural-language interface to this analytical stack. Their strengths include summarization, classification, relation extraction, table generation, and draft scenario construction. Prompting techniques such as role definition, step-by-step reasoning, and chain-of-thought decomposition can improve structured output quality. Retrieval-augmented generation can further ground responses in specific documents and reduce unsupported claims. However, LLMs also exhibit hallucination, inconsistency, sensitivity to prompt wording, and difficulty maintaining formal constraints over long outputs. These weaknesses become critical when generated content enters safety analysis. General AI survey research shows that application value depends on the interaction between model class, data context, and decision objective (Lu,2019a). Reasoning-and-acting research suggests that LLMs perform better when analysis is decomposed into observable intermediate actions (Yao et al.,2023). Large-language-model surveys show that model behavior varies across reasoning, grounding, and alignment settings (Zhao et al.,2023).

The central question is therefore not whether LLMs can produce useful text. They can. The more important question is how LLMs can produce risk-analysis outputs that are constrained enough to be reviewable and computable. Constraint-guided reasoning addresses this question by turning the generation task into a sequence of restricted transformations. The model receives a defined system boundary, a controlled vocabulary of node types, a list of allowed relationships, a set of evidence sources, and a required output schema. The system can then check whether generated nodes and edges satisfy these restrictions. The resulting output is still not automatically true, but it becomes easier for analysts to validate and easier for machines to convert into risk graphs.

Table I. Constraint families for auditable LLM-assisted causal risk analysis.

Constraint family	Purpose	Typical rule	Expected benefit
Structural consistency	Prevents unsupported objects and cross-layer jumps	Nodes and edges must be traceable to supplied system knowledge or approved attack evidence	Reduces invented devices, losses, and causal shortcuts
Task-topology validity	Keeps task failure propagation inside legal process paths	Task-to-task relations must follow documented sequence, message, or gateway logic	Reduces impossible workflow chains
Semantic typing	Prevents concept confusion	Loss, hazard, unsafe control action, task failure, and failure effect must remain distinct	Improves reviewability and model meaning
Evidence traceability	Links each claim to a source	Every node and edge must identify the supporting source class	Improves auditability and accountability
Computable output	Prepares outputs for quantitative modeling	Each generated node must have an identifier, state set, parent list, and evidence field	Improves Bayesian-network mappability

Table I clarifies that constraints should not be treated as optional prompt decorations. They define the boundary between

RESEARCH ARTICLE
Vol. 3, No. 3 (2025), pp. 75–94
DOI: 10.63646/jaiaa.2025.030305

useful AI assistance and uncontrolled narrative generation. In practical deployments, these constraints can be implemented as prompt rules, validation scripts, controlled vocabularies, and reviewer checklists. The stronger the safety requirement, the more important it becomes to preserve this separation between generated candidates and approved risk-model elements. Explainable AI research supports the need to expose the reasoning objects behind each generated causal path (Arrieta et al.,2020). Local explanation methods demonstrate why users need inspectable rationales rather than only final risk labels (Ribeiro et al.,2016).

III. CONSTRAINT-GUIDED LLM REASONING FRAMEWORK

The proposed framework treats an LLM as a constrained translator between heterogeneous industrial knowledge and risk-modeling representation. It does not ask the model to replace safety engineers. Instead, it asks the model to generate candidate causal structures under explicit restrictions so that engineers can review, correct, and approve them. The framework begins with system knowledge: device lists, process descriptions, control objectives, business tasks, alarm logic, operating thresholds, and cyberattack evidence. It then requires the LLM to identify risk objects and relationships in stages. Each stage has a narrower task than the preceding one. This staged design limits the number of assumptions the model needs to make at any one time. Process-mining research supports the use of event and task structures as evidence for workflow-aware risk analysis (van der Aalst et al.,2011). Business-process management research shows that process models can become governance artifacts when they are connected to trusted records (Mendling et al.,2018).

Figure 1 presents the conceptual architecture of the framework. The diagram deliberately avoids arrows to emphasize that the approach is a constrained reasoning stack rather than a simple one-way pipeline. Input knowledge, constraints, LLM reasoning, verification, and risk modeling form mutually reinforcing layers. The system is strongest when all layers are present. If the input knowledge is weak, the model may generate fluent but thin analysis. If the constraints are absent, the model may overgeneralize. If verification is missing, errors can enter the risk model. If the output is not computable, the analysis remains a narrative rather than a decision-support artifact. Digital-twin research similarly frames industrial intelligence as a fusion of data, models, and operational representation (Qi and Tao,2018). Manufacturing digital-twin classifications help distinguish static representation from operationally connected analytical models (Kritzinger et al.,2018). Modeling-oriented digital-twin research clarifies why system representation should be aligned with the decision to be supported (Rasheed et al.,2020).

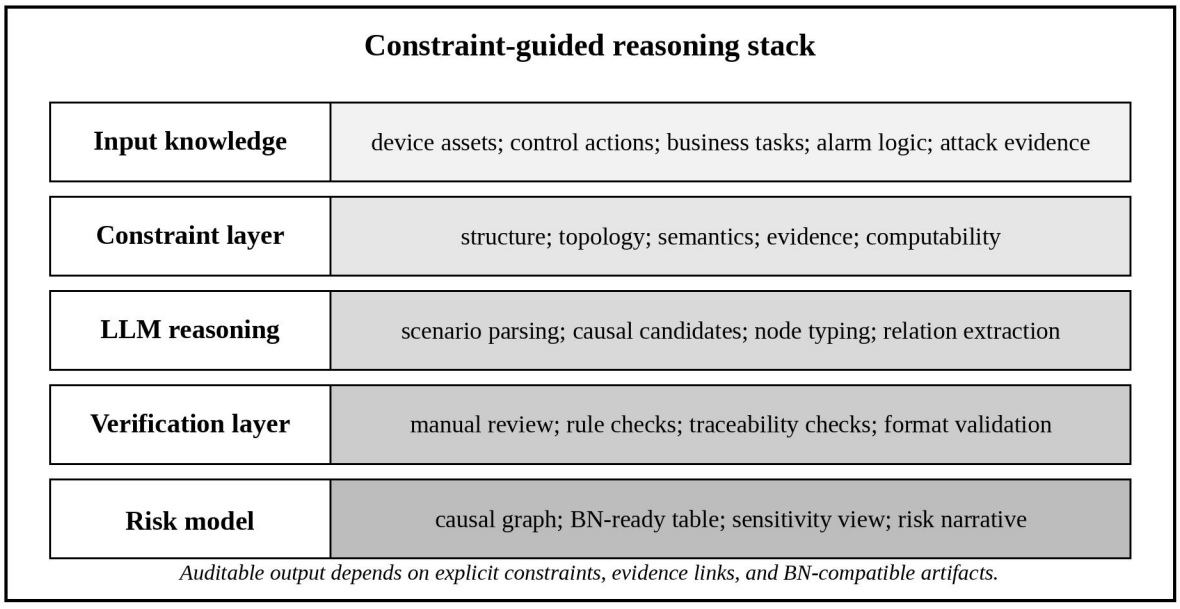

Figure 1. Constraint-guided reasoning stack for causal risk propagation in cyber-physical industrial systems.

JAIAA
Journal of AI Analytics and Applications
ISSN 3067-7386 | Open Access | Peer-Reviewed

RESEARCH ARTICLE
Vol. 3, No. 3 (2025), pp. 75–94
DOI: 10.63646/jaiaa.2025.030305

III.A. Prompt Structure and Analytical Stages

The first stage defines the analysis object. The model receives a boundary statement specifying the industrial system, operational objective, principal process variables, controlled equipment, human roles, and business objectives. This stage prevents the model from expanding the problem into unrelated risks. For example, pressure-control analysis should not generate electricity-market risks unless those risks are part of the specified system boundary. The boundary statement also distinguishes cyberattack evidence from ordinary fault evidence, because the propagation logic may differ between accidental malfunction and adversarial manipulation. Digital-twin surveys identify synchronization, model fidelity, and human use as central challenges for industrial analytics (Fuller et al.,2020).

The second stage identifies typed objects. Instead of asking for a general risk list, the prompt requires separate tables for assets, business tasks, control actions, unsafe control actions, hazards, task failure modes, failure effects, business losses, and safety losses. The model must assign each item to only one primary type. This requirement is essential because many unconstrained outputs confuse causal levels. A failed sensor is not a hazard; it is a possible causal factor. A high-pressure state may be a hazard; it is not itself a cyberattack. Production interruption is a business loss; it is not the same as a delayed alarm task. The typed-object stage creates the vocabulary that later stages must use. Web 3.0 research highlights the wider shift toward decentralized, data-intensive, and trust-sensitive digital infrastructures (Zhang and Lu,2025).

The third stage generates causal candidates. The prompt asks for three kinds of propagation chains: task-centered chains, safety-centered chains, and coupled chains. Task-centered chains describe how task failures produce business effects. Safety-centered chains describe how unsafe control actions and hazardous states produce safety losses. Coupled chains describe the interaction between the task and control perspectives, such as when a missed data-transmission task causes a controller to issue a delayed control action, which then produces a hazardous state and disrupts production. The model is not allowed to introduce new nodes at this stage unless they are added to the typed-object table with evidence. Self-consistency research supports repeated generation and review when the task requires stable reasoning outputs (Wang et al.,2023).

The fourth stage converts causal candidates into a computable table. Each row contains a node identifier, node type, node state, parent nodes, child nodes, evidence source, and reviewer note. This format makes the output suitable for causal graph construction and Bayesian-network modeling. It also creates a clear audit trail: an analyst can trace every generated element back to a process description, control specification, attack hypothesis, or expert assumption. The emphasis on tables may seem less expressive than free-form prose, but it is exactly what makes the output operationally useful.

The fifth stage supports review and calibration. The model is asked to identify ambiguous terms, duplicate nodes, unsupported relationships, and missing evidence links. Human reviewers can then accept, revise, merge, or reject generated items. This final step is essential because constraints improve output quality but do not guarantee correctness. In industrial risk analysis, the model should be treated as an assistant that speeds up candidate generation, not as an authority that validates causal truth. Human-automation interaction theory explains why the proposed workflow keeps analysts in the review loop (Parasuraman et al.,2000). Human-AI interaction guidelines support interface designs that reveal uncertainty, allow correction, and maintain user agency (Amershi et al.,2019).

Table II. Analytical stages in the constraint-guided LLM reasoning workflow.

Stage	Input	LLM task	Reviewer checkpoint
Boundary definition	System scope, assets, process variables, business objectives	Summarize the analysis object and exclude out-of-scope risks	Confirm that system scope and assumptions are complete
Typed-object extraction	Control structure, tasks, attack evidence, loss definitions	Produce separate tables for assets, tasks, UCAs, hazards, failures, and losses	Remove type confusion and duplicate objects
Causal-chain generation	Typed-object tables and allowed relation rules	Generate task-only, safety-only, and coupled propagation chains	Reject unsupported links and causal jumps
Computable mapping	Approved candidate chains	Create node-edge tables with states, parents, children, and sources	Check BN compatibility and acyclicity
Review calibration	Error log from previous stages	Find ambiguous nodes, missing evidence, and inconsistent labels	Accept, revise, merge, or reject candidate elements

The workflow in Table II also supports modular implementation. A team may begin with typed-object extraction and

causal-chain generation, then add computable mapping after a reviewer confirms that the vocabulary is stable. This modularity is important for organizations that already use STPA, FMEA, bow-tie analysis, or process-hazard analysis and want to add LLM assistance without replacing existing governance practices.

III.B. Five Constraint Families

The framework uses five constraint families. Table I summarizes their purpose and expected benefits. Structural consistency is the first constraint because it prevents the model from inventing objects beyond the supplied system knowledge. A model may know that pumps, tanks, boilers, and compressors are common industrial assets, but that general knowledge is not sufficient evidence for a particular analysis. If an item is not in the description of the system, the attack evidence, or an approved reviewer note, it should not enter the causal model. ICS cybersecurity management research emphasizes that risk treatment requires organizational controls as well as technical detection (Knowles et al.,2015).

Task-topology validity is the second constraint. Business processes and control workflows have legal paths. A sent task may precede a receive task; a gateway may split conditions; an alarm may require acknowledgment before a response. Unconstrained models often construct chains that sound reasonable but violate the actual workflow. For example, an operator acknowledgment cannot cause a sensor measurement if the measurement occurs upstream. The topology constraint ensures that task propagation follows the documented process logic.

Semantic typing is the third constraint and is often the most important in practice. Industrial risk analysis uses terms that are related but not interchangeable. A failure mode is a way a task or component fails. The effect of failure is the resulting abnormal state. A hazard is a system state that can lead to loss under worse-case conditions. A loss is an unacceptable consequence. An unsafe control action is a control action that is absent, provided, delayed, early, stopped too soon, or applied too long under a particular context. When these types are mixed, the model becomes difficult to review and may produce invalid quantitative structures. Surveys of black-box explanation methods reinforce the value of making causal artifacts inspectable before they are used in decisions (Guidotti et al.,2018).

Evidence traceability is the fourth constraint. Every node and edge must carry a source label. The source may be a control diagram, process description, task model, attack scenario, historical incident, standard, or expert assumption. Evidence traceability does not prove that a generated causal link is true, but it helps reviewers understand why the link was produced. It also helps distinguish between evidence-based model elements and speculative assumptions. In regulated or safety-critical settings, this distinction is essential. Feature-attribution research further illustrates how explanations can make analytical outputs easier to challenge and calibrate (Lundberg and Lee,2017). Explainability surveys support the requirement that industrial AI outputs be interpretable to reviewers and operators (Adadi and Berrada,2018).

Computable output is the fifth constraint. Free-form risk narratives may be useful for brainstorming, but they cannot be directly used for inference. A Bayesian network requires nodes, states, parent-child relations, and parameters. Even if parameters are estimated later by experts or data, the topology must be unambiguous. The computable-output constraint forces the LLM to produce data structures rather than prose alone. These transforms generated text into an intermediate artifact that can be validated by both humans and software. Guidelines for Bayesian belief-network development emphasize careful node naming, evidence handling, and iterative revision (Marcot et al.,2006).

IV. BENCHMARK DESIGN AND DATA ANALYSIS

To evaluate the framework, this study constructs a synthetic benchmark that is inspired by common industrial cyber-physical settings. The benchmark is not presented as proprietary plant data. It is a controlled evaluation environment designed to test whether constraint-guided LLM reasoning produces more useful causal-risk structures than less constrained prompting strategies. The benchmark includes 24 scenarios distributed across four settings: pressure control in a process unit, water-treatment automation, robotic assembly, and energy microgrid coordination. Each setting includes ordinary fault evidence, cyberattack evidence, task dependencies, control actions, business objectives, and safety objectives. Data-driven smart manufacturing research reinforces the importance of joining sensor streams, production context, and decision analytics (Tao et al.,2018). Manufacturing big-data research demonstrates the operational value of

RESEARCH ARTICLE
Vol. 3, No. 3 (2025), pp. 75–94
DOI: 10.63646/jaiaa.2025.030305

integrating heterogeneous industrial data for proactive action (Wan et al.,2017).

The benchmark design follows three principles. First, each scenario contains cross-domain propagation. The scenario is not limited to an isolated device fault or a generic cyberattack label. It includes at least one device or information abnormality, one task or workflow effect, one control implication, and one business or safety consequence. Second, each scenario includes distractors. These are plausible but out-of-scope items that an unconstrained model might include, such as hazards not linked to the specified process or business losses not supported by the described task chain. Third, each scenario has a reference causal structure prepared from the scenario specification. This reference structure is used to score generated output, not to claim universal ground truth.

Five model configurations are compared. The first uses unconstrained prompting and asks the model to identify causal risk propagation paths. The second uses role-based and step-by-step prompting, instructing the model to act as a safety analyst and reason through the scenario. The third applies the five constraint families without retrieval support. The fourth adds retrieval support by giving the model compact excerpts from the scenario knowledge base before each generation step. The fifth adds review calibration, meaning that the model receives a short error log from the previous output and is asked to revise the table under the same constraints. This progression allows the analysis to separate the value of constraints from the value of retrieval and iterative review. Digital twin-driven smart manufacturing research shows that reference models are useful only when they support operational inference and action (Lu et al.,2020). AI-driven digital-twin surveys connect industrial AI with system representation, monitoring, and cyber-physical feedback (Huang et al.,2021).

The evaluation uses five metrics. Causal validity measures the share of generated relationships that match the reference structure or are judged admissible extensions. Semantic precision measures whether generated nodes are assigned to the correct type. Bayesian network capability measures whether output rows contain enough structured information to become nodes and edges in a probabilistic model. Redundancy measures the share of generated items that duplicate other items under different wording. Hallucination rate measures the share of items that are unsupported by supplied scenario evidence. Because the goal is not only accuracy but also usability, these metrics jointly evaluate engineering quality, auditability, and computability.

Table III summarizes the benchmark scenario families. Each scenario family was designed to represent a different industrial logic. Pressure control emphasizes continuous process dynamics and safety limits. Water treatment emphasizes regulatory service continuity and delayed quality effects. Robotic assembly emphasizes discreet-event coordination, human proximity, and motion safety. Energy microgrid coordination emphasizes distributed control, load balancing, and service reliability. These settings are diverse enough to test whether the constraint framework is tied to one industrial domain or generalizable across cyber-physical systems.

Table III. Synthetic benchmark scenario families used for evaluation.

Scenario family	Primary cyber-physical disturbance	Risk-propagation focus	Typical consequence represented
Pressure control unit	False pressure reading, delayed actuator response, alarm suppression	Sensor-task-control coupling and unsafe pressure state	Unsafe operating state, shutdown, product-quality deviation
Water-treatment automation	Manipulated chlorine sensor data, delayed pump task, compromised supervisory setpoint	Quality monitoring, safety interlock, and service continuity	Regulatory noncompliance, service interruption, health-related risk
Robotic assembly cell	Unauthorized speed parameter change, delayed emergency-stop signal, camera spoofing	Human-machine proximity and discrete task sequencing	Unsafe robot motion, scrap, operator exposure
Energy microgrid coordination	Compromised load forecast, false inverter status, delayed dispatch command	Distributed control, load prioritization, and resilience	Service instability, load shedding, equipment stress

The benchmark does not attempt to cover every industrial risk category. Its purpose is to test reasoning quality under conditions that are common across industrial systems: incomplete evidence, cross-domain dependencies, time-sensitive control, and business consequences. These conditions are exactly where unconstrained LLM outputs tend to become verbose, speculative, or difficult to map into a formal model. Information-systems research on blockchain implementation also underscores the need to link technical design with process governance (Lu,2022).

RESEARCH ARTICLE
Vol. 3, No. 3 (2025), pp. 75–94
DOI: 10.63646/jaiaa.2025.030305

V. RESULTS

The benchmark results indicate that constraints make the largest difference when the output must support downstream computation. Unconstrained prompting generated long and fluent explanations, but many outputs included invented intermediate nodes, combined hazards with losses, or jumped directly from attack evidence to final consequences. Role-based and step-by-step prompting improved organization but did not reliably enforce legal causal types. Constraint-guided prompting produced shorter outputs, but those outputs were substantially easier to review and map into causal graphs. This finding is important because risk analysts often do not need more text; they need cleaner structure.

Figure 2 compares the five model configurations across causal validity, Bayesian network capability, and hallucination control. Causal validity improves from 0.58 under unconstrained prompting to 0.82 under constraint-guided prompting. Bayesian network capability improves from 0.43 to 0.76. Hallucination control, represented as the inverse of hallucination rate, improves from 0.71 to 0.88. Retrieval and review calibration add further gains, with the review-calibrated configuration reaching 0.92 causal validity and 0.84 Bayesian network capability. The improvement is not merely cosmetics. The output generated becomes better aligned with the data structures needed for quantitative inference.

The cycle review analysis in Figure 3 provides a second perspective. Accepted causal chains increased steadily over five review cycles, while redundant chains and missing evidence links declined. This pattern suggests that LLM-assisted analysis can benefit from iterative calibration when the feedback is structured. General feedback such as “be more precise” is less useful than specific feedback such as “do not label device failures as hazards” or “each edge must include a source.” The most effective review cycle is therefore not a conversation in ordinary language, but a controlled correction loop tied to the same constraint families used during generation.

Table IV reports the main quantitative results. The largest relative gain appears in Bayesian network capability, which increases by 0.33 points between unconstrained and constraint-guided prompting. This result is expected because the computable-output constraint directly targets the information required for causal graph construction. The second-largest gain appears in semantic precision, indicating that explicit type rules reduce confusion among failures, hazards, unsafe control actions, and losses. Hallucination rate falls from 28.6% to 12.4% under the constraint-guided configuration and to 7.3% after review calibration. Redundancy falls as well, which reduces reviewer burden. Digital-twin characterization research emphasizes that a useful twin must have a clear purpose, data relation, and lifecycle role (Jones et al.,2020).

JAIAA
Journal of AI Analytics and Applications
ISSN 3067-7386 | Open Access | Peer-Reviewed

RESEARCH ARTICLE
Vol. 3, No. 3 (2025), pp. 75–94
DOI: 10.63646/jaiaa.2025.030305

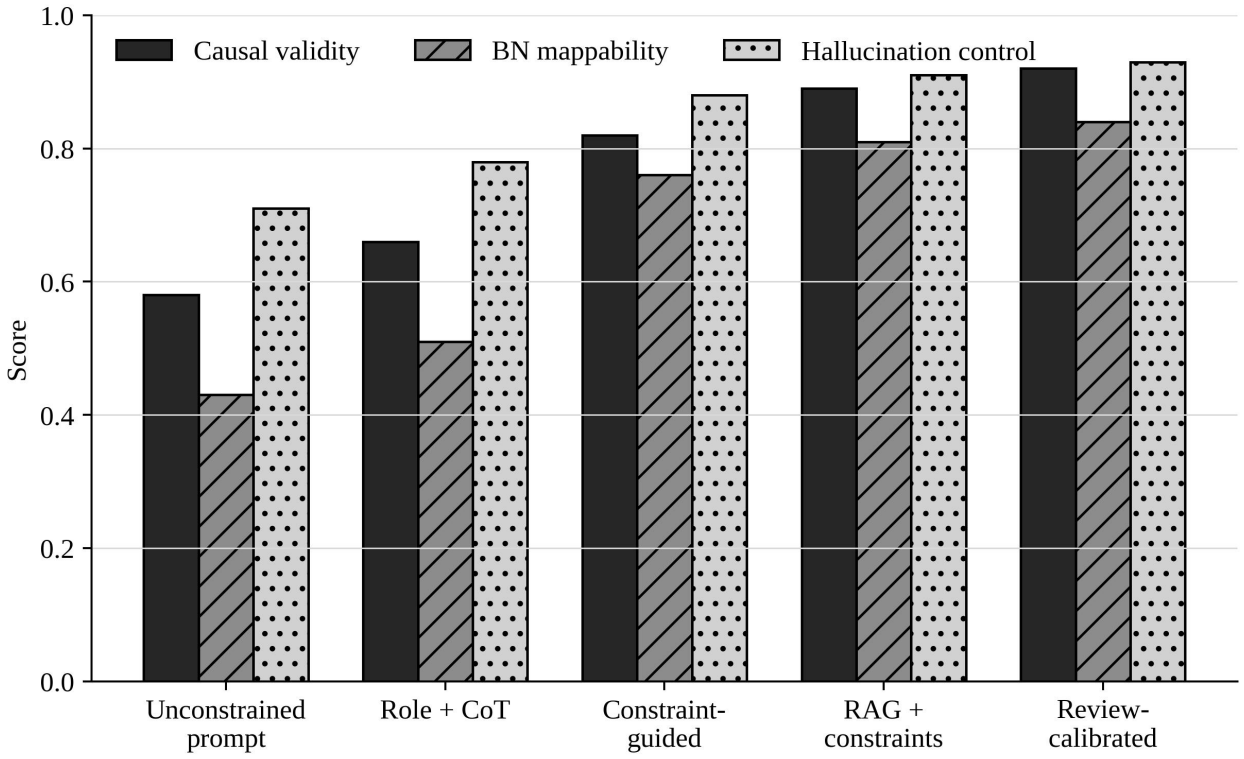

Figure 2. Benchmark comparison of LLM prompting configurations across causal validity, BN mappability, and hallucination control.

Table IV. Quantitative benchmark results across LLM reasoning configurations.

Configuration	Causal validity	Semantic precision	BN mappability	Hallucination rate	Redundancy rate
Unconstrained prompt	0.58	0.61	0.43	28.6%	24.8%
Role + step-by-step reasoning	0.66	0.69	0.51	22.1%	20.3%
Constraint-guided reasoning	0.82	0.86	0.76	12.4%	13.7%
RAG + constraints	0.89	0.90	0.81	9.1%	11.5%
Review-calibrated constraints	0.92	0.93	0.84	7.3%	9.6%

The results in Table IV show that the constraint-guided configuration offers the strongest single-step improvement. Retrieval and review calibration improve the outcome further, but they operate on top of the structural discipline imposed by the constraints. In other words, retrieval provides better evidence, and review calibration provides better correction, but constraints provide the grammar of acceptable risk modeling. Communication-system survey research also suggests that future industrial systems will rely on increasingly distributed and low-latency connectivity (Lu and Zheng,2020).

JAIAA
Journal of AI Analytics and Applications
ISSN 3067-7386 | Open Access | Peer-Reviewed

RESEARCH ARTICLE
Vol. 3, No. 3 (2025), pp. 75–94
DOI: 10.63646/jaiaa.2025.030305

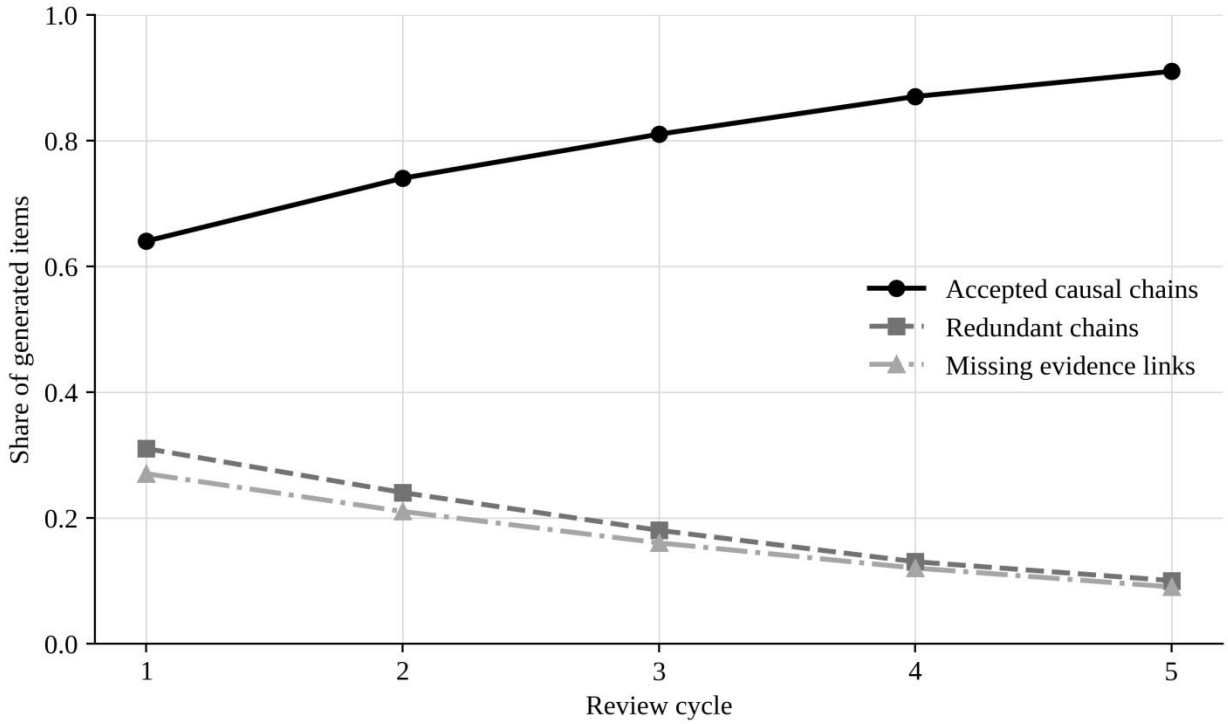

Figure 3. Review-cycle dynamics for accepted causal chains, redundancy, and missing evidence links.

V.A. Ablation Analysis of Constraint Families Reviews of quantum machine learning show that advanced analytical methods still require classification of use cases, constraints, and implementation limits (Lu et al.,2024a).

The ablation analysis examines which constraint families reduce which error types. Figure 4 presents an estimated impact matrix. Structural consistency has the strongest effect on invented nodes and illegal cross-layer jumps. Task-topology validity has the strongest effect on illegal edges inside business-process chains. Semantic typing has the strongest effect on type confusion. Evidence traceability has the strongest effect on unsupported claims. Computable output has the strongest effect on missing Bayesian-network table fields. The matrix confirms that no single constraint is sufficient. Each constraint protects a different part of the risk-analysis workflow.

The ablation results also clarify why ordinary prompt engineering is insufficient. A role prompt may remind the model to behave like a safety engineer, but it does not specify which edges are legal. A chain-of-thought prompt may encourage stepwise reasoning, but it does not prevent the model from creating unsupported intermediate variables. A format prompt may produce a table, but it does not ensure that table rows are semantically valid. Constraint-guided reasoning works because it combines role, reasoning, evidence, and output restrictions in one integrated workflow.

The most persistent remaining error is under-specification. In some scenarios, the model followed all formal constraints but produced causal chains that were too general. For example, it might generate “delayed control response causes process instability” without specifying the affected process variable, threshold condition, or timing context. This output is not necessarily wrong, but it is less useful for quantitative inference. The finding suggests that future systems should include parameter prompts that request threshold ranges, time windows, and operational contexts whenever such values are available.

RESEARCH ARTICLE
Vol. 3, No. 3 (2025), pp. 75–94
DOI: 10.63646/jaiaa.2025.030305

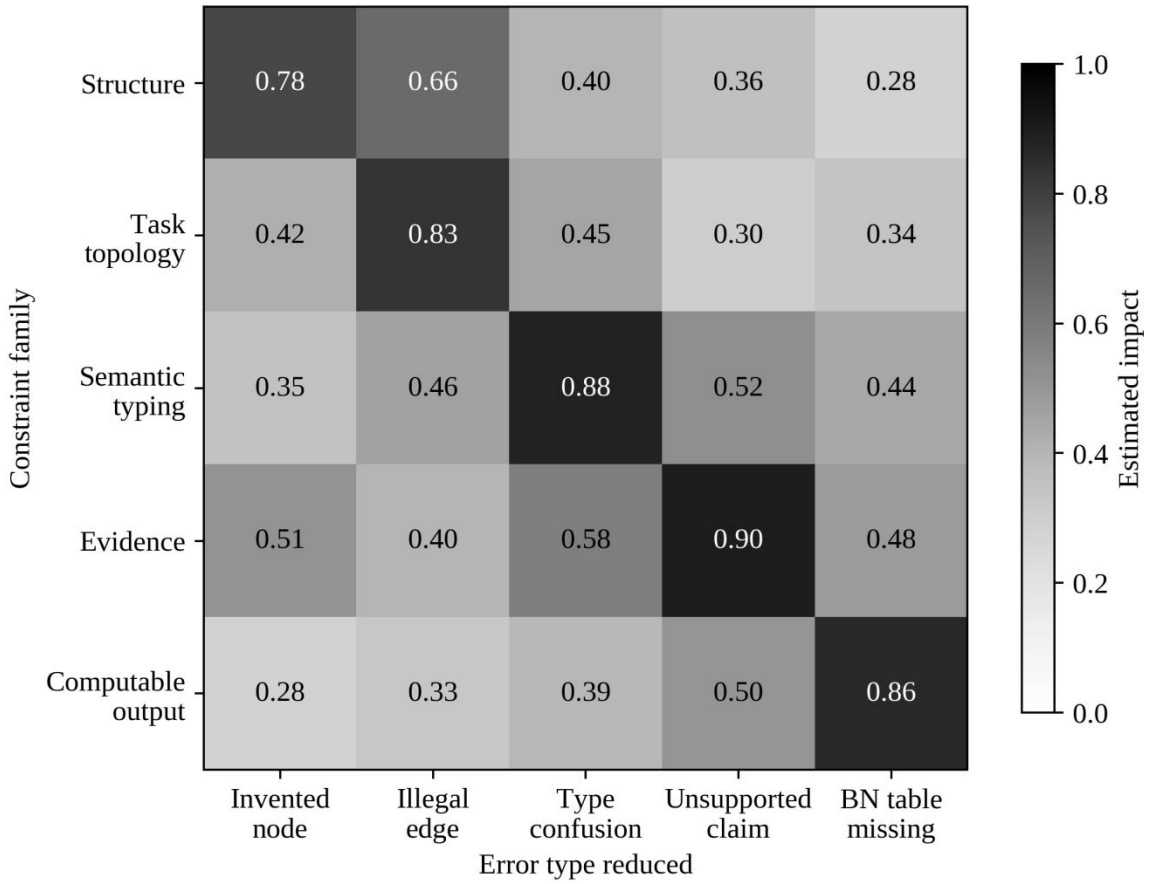

Figure 4. Estimated impact of constraint families on common LLM risk-modeling errors.

Table V. Ablation results when individual constraint families are removed.

Removed constraint	Most affected metric	Observed degradation	Interpretation
Structural consistency	Hallucination rate	+8.9 percentage points	The model created plausible but unsupported device and state nodes
Task-topology validity	Causal validity	-0.11	The model connected tasks that were not adjacent or legally linked
Semantic typing	Semantic precision	-0.18	The model confused hazards, losses, UCAs, and task effects
Evidence traceability	Audit completeness	-0.24	Reviewers could not determine why several edges were generated
Computable output	BN mappability	-0.29	The output lacked states, parents, and identifiers required for inference

Table V reinforces the interpretation of Figure 4. Removing any single constraint weakens the workflow, but the type of degradation differs. The strongest operational lesson is that semantic typing and computable output should not be postponed until the end of the analysis. If type confusion and missing identifiers are allowed to accumulate, later review becomes slower and more error prone. Management analytics research frames data-driven decision support as an interdisciplinary field linking technology, organization, and analysis.

V.B. Scenario-Level Risk Analysis

The final part of the evaluation tests whether cleaner causal structures lead to more interpretable scenario-level risk analysis. Each benchmark scenario was converted into a normalized risk index from 0 to 1 after reviewer approval. The risk index is not intended to replace plant-specific quantitative safety assessment. It is a comparative score that reflects the probability-weighted severity of safety and business consequences under fault-only evidence, attack evidence, and constraint-guided mitigation assumptions. The mitigation assumptions represent responses such as added evidence checks,

RESEARCH ARTICLE
Vol. 3, No. 3 (2025), pp. 75–94
DOI: 10.63646/jaiaa.2025.030305

alarm validation, redundant task confirmation, or safer default states. Work on stealthy industrial attacks shows that seemingly small manipulation can create disproportionate effects when process dynamics are tightly coupled (Urbina et al.,2016). Physics-based detection research highlights the importance of modeling process behavior rather than only network events (Giraldo et al.,2018). Digital supply-chain resilience research emphasizes end-to-end visibility, which is also essential for tracing cyber-physical risk propagation (Ivanov,2021).

Figure 5 shows that attack evidence substantially increases the risk index in all four scenario families. The pressure-control unit rises from 0.42 to 0.71, reflecting the severity of false pressure readings and delayed actuator response. Water-treatment automation rises from 0.36 to 0.64 because manipulated quality measurements can delay regulatory intervention. Robotic assembly rises from 0.31 to 0.56 because unauthorized motion parameters can create human-proximity hazards. Energy microgrid coordination rises from 0.39 to 0.68 because false inverter status can disturb distributed load balancing. Under mitigation assumptions informed by constraint-guided causal analysis, risk falls toward the fault-only level but does not return fully to baseline. Management analytics research further supports the paper's emphasis on converting analytical outputs into decision-relevant artifacts (Lu et al.,2024b).

This result has two implications. First, constraint-guided LLM reasoning is useful not only for scenario generation but also for identifying where controls should be placed. If a generated chain clearly links false data to a task failure and then to an unsafe control action, the analyst can consider evidence validation at the data boundary, timeout logic at the task boundary, or fallback control at the control-action boundary. Second, the gap between mitigated risk and baseline risk reminds us that cyberattack evidence changes the system context. Even a well-designed mitigation may not erase increased uncertainty; it may only reduce it to a tolerable level.

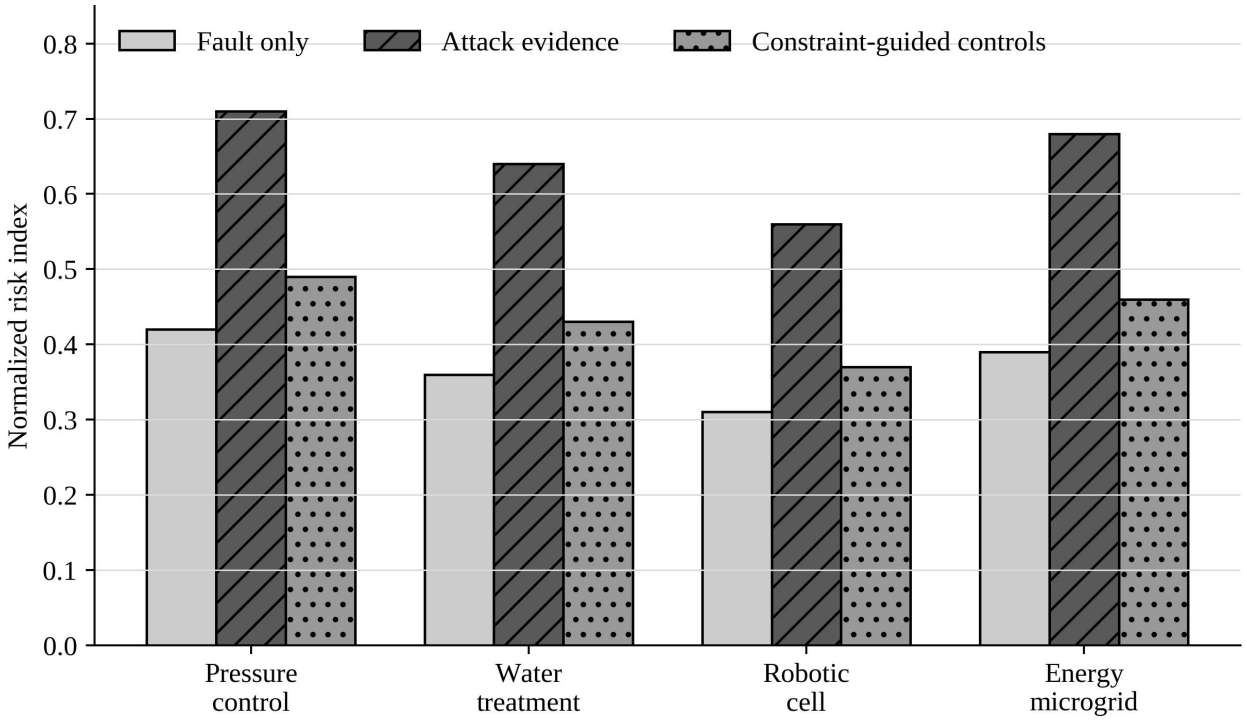

Figure 5. Scenario-level normalized risk index under fault-only evidence, attack evidence, and constraint-guided mitigation assumptions.

VI. DISCUSSION

The results support the central argument: LLMs are most valuable in industrial risk analysis when they are constrained. The reason is not that constraints make the model more intelligent in a general sense. Rather, constraints make the model more useful for a specific engineering workflow. In free-form mode, the model behaves like a fluent analyst who may

RESEARCH ARTICLE
Vol. 3, No. 3 (2025), pp. 75–94
DOI: 10.63646/jaiaa.2025.030305

overreach. In constraint-guided mode, it behaves more like a table-preparation assistant that must respect a controlled vocabulary, documented topology, and evidence structure. The latter role is less dramatic but far more compatible with safety-critical analysis. Critical-infrastructure cybersecurity research reinforces the need to connect technical events with organizational and production consequences (Ani et al.,2017).

The framework also changes how organizations should think about AI trust. Trust should not be placed in the LLM alone. It should be placed in the socio-technical workflow that contains the LLM. That workflow includes source documents, prompts, constraints, validation rules, reviewer expertise, and downstream quantitative checks. A generated causal link is trustworthy only after it has survived this workflow. This perspective aligns with broader arguments in interpretable machine learning and AI governance: explanations and audit logs matter because they provide the context in which humans can challenge model outputs. Foundation-model research emphasizes that broad capability must be balanced with governance, evaluation, and domain grounding (Bommasani et al.,2021). High-stakes interpretability arguments support the preference for structured causal tables over opaque generated narratives (Rudin,2019).

The findings also suggest that industrial LLM systems should be designed for intermediate artifacts rather than final answers. A final answer such as “the system risk is high” is too compressed to be audited. An intermediate artifact such as a node-edge table can be checked, corrected, parameterized, and reused. It can also be versioned. When a plant changes a controller, task sequence, or operating threshold, analysts can revise the affected nodes and rerun inference rather than rewriting a narrative from scratch. This artifact-oriented design is especially relevant for industrial settings where procedures, equipment, and attack surfaces evolve over time. Industry 4.0 research also identifies integration, analytics, and open system design as core issues in smart industrial transformation (Lu,2017b).

A practical concern is the cost of preparing system knowledge. Constraint-guided reasoning requires structured descriptions of tasks, control actions, assets, and evidence sources. Many organizations do not maintain such descriptions in a clean digital form. However, this requirement should be viewed as a benefit rather than a weakness. If an organization cannot describe its task topology, control boundaries, and evidence sources, then its risk assessment is already fragile. The LLM framework exposes the need for better knowledge management. It also provides an incentive to build reusable libraries of assets, task types, control actions, hazards, and evidence classes. Research on digital twin enabling technologies shows that integration depends on data pipelines, modeling tools, and interoperable infrastructures (Qi et al.,2021). Blockchain research in industrial information integration highlights traceability as a recurring requirement for distributed decision environments (Lu,2019b).

Another concern is whether constraints will make the model too conservative and miss novel risk paths. This is possible. A strict structural constraint may prevent the model from proposing a previously unrecognized dependency. The solution is not to remove constraints but to separate exploratory generation from approved modeling. During exploratory analysis, the model may be allowed to propose “candidate new dependencies” in a separate table marked as speculative. While approved modeling, only reviewed and evidence-supported items enter the causal graph. This separation preserves creative discovery while protecting the quantitative model from unsupported speculation. Emerging research on quantum information systems points to a broader trend toward intelligent algorithms that operate across multiple industrial domains (Wang and Lu,2026).

The framework is also compatible with retrieval-augmented generation. Retrieval is useful because it narrows the evidence space. Instead of relying on general model memory, the LLM receives targeted excerpts from process descriptions, alarm rationales, task models, and incident records. Yet retrieval alone does not solve semantic confusion or output computability. A retrieved sentence may mention a pump failure and a safety shutdown, but the model still needs constraints to represent which item is a causal factor, which item is a control action, which item is a hazard, and which item is a loss. Retrieval and constraints should therefore be treated as complementary design elements. Retrieval-augmented generation provides a technical route for tying LLM outputs to external evidence rather than model memory alone (Lewis et al.,2020). Recent RAG surveys show that retrieval quality, reranking, and answer control must be evaluated as a combined system (Gao et al.,2024). Dense retrieval research provides one technical foundation for retrieving relevant industrial evidence before generation (Karpukhin et al.,2020). Retrieval-enhanced language modeling

RESEARCH ARTICLE
Vol. 3, No. 3 (2025), pp. 75–94
DOI: 10.63646/jaiaa.2025.030305

research indicates that external evidence can improve factual coverage when retrieval is well controlled (Borgeaud et al.,2022).

The discussion would be incomplete without limitations. The benchmark is synthetic, even though it is domain informed. Real industrial systems include messy documents, missing diagrams, inconsistent labels, proprietary controller logic, and organizational practices that cannot be fully represented in a scenario specification. The evaluation metrics depend on reference structures prepared from scenario descriptions, and different experts may disagree on admissible causal links. The study also does not estimate conditional probability tables from real plant data. Instead, it focuses on the quality of causal structures that could later be parameterized. Future research should test the framework in real industrial case studies, compare expert review time, and evaluate whether constraint-guided LLM outputs improve the accuracy and usefulness of downstream Bayesian inference. Deep-learning research in smart manufacturing shows why model performance must be evaluated against manufacturing-specific data and process requirements (Wang et al.,2018).

VII. IMPLEMENTATION GUIDELINES

Organizations seeking to apply constraint-guided LLM reasoning should begin with a narrow use case. A full plant-level model is too broad for first deployment. A better starting point is one critical control loop, one safety function, one business workflow, or one high-priority attack scenario. The first goal should be to produce a reviewed node-edge table, not a complete enterprise risk model. Once the team understands how constraints operate, the method can be extended to additional loops and workflows. Recent Industry 4.0 trend analysis further highlights the continuing shift toward integrated, intelligent, and data-intensive industrial systems (Lu,2025).

The second guideline is to create a controlled vocabulary before prompting. The vocabulary should include node types, allowed relation types, task names, device names, control actions, and loss categories. If the vocabulary is missing, the model will invent terminology. If terminology differs across documents, the model should be asked to map synonyms before causal-chain generation. This step reduces duplicate nodes and improves the stability of later Bayesian-network construction. Research on blockchain trends highlights the long-term importance of trusted coordination mechanisms in enterprise systems (Zheng and Lu,2022).

The third guideline is to preserve evidence links. Each generated item should carry a short evidence label. The evidence label may point to a diagram, procedure, alarm specification, attack hypothesis, data record, or expert assumption. This does not need to be complex in the first version. Even a simple source class is better than no source at all. Evidence links allow reviewers to ask the most important question in risk analysis: “Why is this element in the model?” Research on blockchain-enabled auditing illustrates why traceable records matter for accountability in analytical workflows (Wu et al.,2025).

The fourth guideline is to design reviews as a structured workflow. Reviewers should not merely read the generated text. They should accept, reject, merge, split, or revise each node and edge. The review log should identify error categories. These categories can then be fed back into the next LLM run. Over time, the organization builds a reusable error taxonomy that improves consistency across analysts and systems. Studies of interpretability practice show that explanation tools must fit expert workflows rather than simply generate more information (Kaur et al.,2020).

The fifth guideline is to separate causal modeling from probability estimation. LLMs may assist in preparing a topology, but conditional probabilities should be estimated through historical data, expert elicitation, simulation, or calibrated assumptions. Mixing generated causal text with unsupported numerical probabilities can create false precision. A safe workflow first approves the graph structure and then parameterizes it through an explicit probability procedure. Ranked-node Bayesian modeling provides a practical way to convert qualitative expert judgments into probabilistic structures (Fenton et al.,2007). Bayesian-network applications in complex modeling domains show how uncertainty can be represented when causal knowledge is incomplete (Aguilera et al.,2011).

The final guideline is to treat the LLM system as part of the safety-management process. Prompt templates, constraint rules, review logs, and approved outputs should be versioned. Changes to equipment, procedures, or cyber controls should

RESEARCH ARTICLE
Vol. 3, No. 3 (2025), pp. 75–94
DOI: 10.63646/jaiaa.2025.030305

trigger model review. The organization should define who is authorized to approve the risk elements generated and how disagreements are resolved. These governance details may appear administrative, but they determine whether AI-assisted risk analysis becomes a reliable engineering practice or a collection of isolated experiments. Earlier blockchain analytics research also highlights the need to examine security, trust, and integration together (Lu,2018).

Table VI. Practical implementation guidelines for constraint-guided LLM risk reasoning.

Implementation step	Recommended practice	Common failure to avoid
Scope the use case	Begin with one control loop, task chain, or attack scenario	Attempting a full plant model before terms and review rules are stable
Build vocabulary	Define approved node types, relation types, task names, and loss categories	Allowing the model to invent labels that duplicate existing concepts
Prompt with constraints	Use explicit rules for structure, topology, semantics, evidence, and output schema	Using only a role prompt such as “act as a safety expert”
Review systematically	Accept, reject, merge, split, or revise each generated node and edge	Reviewing only the narrative conclusion
Parameterize separately	Estimate probabilities using data, expert elicitation, or simulation	Treating generated probabilities as verified evidence
Version the workflow	Store prompts, sources, outputs, reviewer decisions, and model versions	Losing traceability when the system or prompt changes

The guidelines in Table VI are intentionally operational. They translate the conceptual framework into a sequence of actions that a safety, cybersecurity, or industrial analytics team can implement. The most important message is that the LLM should be surrounded by stronger process discipline than an ordinary drafting tool because its output may affect technical risk priorities. Industrial information-integration research on emerging computing paradigms illustrates how analytical infrastructures are expanding beyond conventional software stacks (Lu et al.,2023).

VIII. CONCLUSION

Cyber-physical industrial systems require risk-analysis methods that connect devices, tasks, control actions, hazards, losses, and business objectives. LLMs can assist this work because they can transform heterogeneous descriptions into structured candidate causal models. Yet unconstrained LLM generation is not reliable enough for safety-critical use. It can create unsupported nodes, confuse risk concepts, violate task topology, and generate outputs that cannot be translated into quantitative models.

This paper developed a constraint-guided LLM reasoning framework for causal risk propagation in industrial cyber-physical systems. The framework uses five constraint families: structural consistency, task-topology validity, semantic typing, evidence traceability, and computable output formatting. A benchmark study across 24 synthetic scenarios showed that constraint-guided reasoning substantially improves causal validity, semantic precision, Bayesian network capability, and hallucination control compared with unconstrained prompting and role-based prompting. Retrieval support and review calibration provide additional gains, but their value depends on the underlying constraint structure.

The broader contribution is methodological. The paper shows how LLMs can be integrated into industrial AI analytics without treating generated language as final evidence. The appropriate role of an LLM is to accelerate the preparation of auditable intermediate artifacts that humans and risk models can examine. Future work should test the framework on real plant data, incorporate multimodal sources such as diagrams and event logs, and connect constraint-guided causal modeling with dynamic Bayesian inference, digital twins, and operational decision support. As industrial systems become more connected and more exposed to cyber-physical disruption, the ability to reason about causal risk propagation in an auditable way will become a central requirement for responsible industrial AI.

AUTHOR CONTRIBUTIONS

Author	Contribution
Marta Fabbri	Conceptualization, methodology, writing - original draft, visualization
Giorgio Bianchi	Formal analysis, benchmark design, data curation, validation
Elena Greco	Supervision, writing - review and editing, project administration, correspondence

DECLARATIONS

Conflicts of interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this manuscript.

Data availability: The benchmark data used in this manuscript are synthetic and generated from scenario specifications described in the paper. No proprietary industrial dataset or confidential plant record is redistributed.

Funding: This research received no external funding.

Ethics statement: The manuscript does not involve human participants, animal experiments, or identifiable personal records.

ABOUT THE AUTHORS

Marta Fabbri is affiliated with the Department of Information Engineering, University of Brescia, Italy. Her research focuses on industrial AI, cyber-physical system monitoring, and interpretable analytics for manufacturing environments.

Giorgio Bianchi is affiliated with the Department of Industrial Engineering, University of Salerno, Italy. His research interests include industrial safety analysis, process control, and cyber-resilient production systems.

Elena Greco is affiliated with the Department of Computer Science, University of Bari Aldo Moro, Italy. Her research addresses knowledge-guided AI, causal reasoning, and trustworthy decision support for cyber-physical systems.

REFERENCES

Lee, J., Bagheri, B., & Kao, H.-A. (2015). A cyber-physical systems architecture for Industry 4.0-based manufacturing systems. Manufacturing Letters, 3, 18–23. https://doi.org/10.1016/j.mfglet.2014.12.001

Yang, L., Hou, Q., Zhu, X., Lu, Y., & Xu, L. D. (2025). Potential of large language models in blockchain-based supply chain finance. Enterprise Information Systems, 19(11), 2541199. https://doi.org/10.1080/17517575.2025.2541199

Humayed, A., Lin, J., Li, F., & Luo, B. (2017). Cyber-physical systems security: A survey. IEEE Internet of Things Journal, 4(6), 1802–1831. https://doi.org/10.1109/JIOT.2017.2703172

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys, 55(9), Article 195. https://doi.org/10.1145/3560815

Lu, Y. (2017a). Cyber physical system (CPS)-based Industry 4.0: A survey. Journal of Industrial Integration and Management, 2(3), 1750014. https://doi.org/10.1142/S2424862217500142

Cherdantseva, Y., Burnap, P., Blyth, A., Eden, P., Jones, K., Soulsby, H., & Stoddart, K. (2016). A review of cyber security risk assessment methods for SCADA systems. Computers & Security, 56, 1–27. https://doi.org/10.1016/j.cose.2015.09.009

Zhang, C., & Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. Journal of Industrial Information Integration, 23, 100224. https://doi.org/10.1016/j.jii.2021.100224

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30. https://doi.org/10.48550/arXiv.1706.03762

Lu, Y., & Xu, L. D. (2019). Internet of Things cybersecurity research: A review of current research topics. IEEE Internet of Things Journal, 6(2), 2103–2115. https://doi.org/10.1109/JIOT.2018.2869847

Tao, F., Qi, Q., Liu, A., & Kusiak, A. (2018). Data-driven smart manufacturing. Journal of Manufacturing Systems, 48, 157–169. https://doi.org/10.1016/j.jmsy.2018.01.006

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623. https://doi.org/10.1145/3442188.3445922

Pasqualetti, F., Dörfler, F., & Bullo, F. (2013). Attack detection and identification in cyber-physical systems. IEEE Transactions on Automatic Control, 58(11), 2715–2729. https://doi.org/10.1109/TAC.2013.2266831

Scutari, M. (2010). Learning Bayesian networks with the bnlearn R package. Journal of Statistical Software, 35(3), 1–22. https://doi.org/10.18637/jss.v035.i03

Sisinni, E., Saifullah, A., Han, S., Jennehag, U., & Gidlund, M. (2018). Industrial Internet of Things: Challenges, opportunities, and directions. IEEE Transactions on Industrial Informatics, 14(11), 4724–4734. https://doi.org/10.1109/TII.2018.2852491

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits

reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824–24837. https://doi.org/10.48550/arXiv.2201.11903

Bowles, J. B., & Peláez, C. E. (1995). Fuzzy logic prioritization of failures in a system failure mode, effects and criticality analysis. Reliability Engineering & System Safety, 50(2), 203–213. https://doi.org/10.1016/0951-8320(95)00068-D

Qi, Q., & Tao, F. (2018). Digital twin and big data towards smart manufacturing and Industry 4.0: 360 degree comparison. IEEE Access, 6, 3585–3593. https://doi.org/10.1109/ACCESS.2018.2793265

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901. https://doi.org/10.48550/arXiv.2005.14165

Xu, L. D., Lu, Y., & Li, L. (2021). Embedding blockchain technology into IoT for security: A survey. IEEE Internet of Things Journal, 8(13), 10452–10473. https://doi.org/10.1109/JIOT.2021.3060508

Kriaa, S., Pietre-Cambacedes, L., Bouissou, M., & Halgand, Y. (2015). A survey of approaches combining safety and security for industrial control systems. Reliability Engineering & System Safety, 139, 156–178. https://doi.org/10.1016/j.ress.2015.02.008

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & 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

Lu, Y. (2019a). Artificial intelligence: A survey on evolution, models, applications and future trends. Journal of Management Analytics, 6(1), 1–29. https://doi.org/10.1080/23270012.2019.1570365

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. Machine Learning, 29(2–3), 131–163. https://doi.org/10.1023/A:1007465528199

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Chin, K.-S., Wang, Y.-M., Poon, G. K. K., & Yang, J.-B. (2009). Failure mode and effects analysis using a group-based evidential reasoning approach. Computers & Operations Research, 36(6), 1768–1779. https://doi.org/10.1016/j.cor.2008.05.002

Zhang, H., & Lu, Y. (2025). Web 3.0: Applications, opportunities and challenges in the next internet generation. Systems Research and Behavioral Science, 42(4), 996–1015. https://doi.org/10.1002/sres.3151

Urbina, D. I., Giraldo, J. A., Cárdenas, A. A., Tippenhauer, N. O., Valente, J., Faisman, B., Ruth, B., Candell, R., & Sandberg, H. (2016). Limiting the impact of stealthy attacks on industrial control systems. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 1092–1105. https://doi.org/10.1145/2976749.2978388

Kritzinger, W., Karner, M., Traar, G., Henjes, J., & Sihn, W. (2018). Digital twin in manufacturing: A categorical literature review and classification. IFAC-PapersOnLine, 51(11), 1016–1022. https://doi.org/10.1016/j.ifacol.2018.08.474

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2210.03629

Chen, Y., Lu, Y., Bulysheva, L., & Kataev, M. Y. (2024). Applications of blockchain in Industry 4.0: A review. Information Systems Frontiers, 26(5), 1715–1729. https://doi.org/10.1007/s10796-022-10248-7

Liu, H.-C., Liu, L., & Liu, N. (2013). Risk evaluation approaches in failure mode and effects analysis: A literature review. Expert Systems with Applications, 40(2), 828–838. https://doi.org/10.1016/j.eswa.2012.08.010

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778

Lu, Y. (2022). Implementing blockchain in information systems: A review. Enterprise Information Systems, 16(12), 1876–1907. https://doi.org/10.1080/17517575.2021.2008513

Fuller, A., Fan, Z., Day, C., & Barlow, C. (2020). Digital twin: Enabling technology, challenges and open research. IEEE Access, 8, 108952–108971. https://doi.org/10.1109/ACCESS.2020.2998358

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), Article 248. https://doi.org/10.1145/3571730

Marcot, B. G., Steventon, J. D., Sutherland, G. D., & McCann, R. K. (2006). Guidelines for developing and updating Bayesian belief networks. Canadian Journal of Forest Research, 36(12), 3063–3074. https://doi.org/10.1139/x06-135

Lu, Y. (2019b). The blockchain: State-of-the-art and research challenges. Journal of Industrial Information Integration, 15, 80–90. https://doi.org/10.1016/j.jii.2019.04.002

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2203.11171

Rasheed, A., San, O., & Kvamsdal, T. (2020). Digital twin: Values, challenges and enablers from a modeling perspective. IEEE Access, 8, 21980–

22012. https://doi.org/10.1109/ACCESS.2020.2970143

Teixeira, A., Shames, I., Sandberg, H., & Johansson, K. H. (2015). A secure control framework for resource-limited adversaries. Automatica, 51, 135–148. https://doi.org/10.1016/j.automatica.2014.10.067

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30. https://doi.org/10.48550/arXiv.1705.07874

Zheng, X. R., & Lu, Y. (2022). Blockchain technology: Recent research and future trend. Enterprise Information Systems, 16(12), 1939895. https://doi.org/10.1080/17517575.2021.1939895

van der Aalst, W. M. P., Adriansyah, A., de Medeiros, A. K. A., Arcieri, F., Baier, T., Blickle, T., Bose, J. C., van den Brand, P., Brandtjen, R., Buijs, J., Burattin, A., Carmona, J., Castellanos, M., Claes, J., Cook, J., Costantini, N., Curbera, F., Dongen, B. F., ... Wynn, M. (2011). Process mining manifesto. Business Process Management Workshops, 99, 169–194. https://doi.org/10.1007/978-3-642-28108-2_19

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. arXiv. https://doi.org/10.48550/arXiv.2312.10997

Weber, P., Medina-Oliva, G., Simon, C., & Iung, B. (2012). Overview on Bayesian networks applications for dependability, risk analysis and maintenance areas. Engineering Applications of Artificial Intelligence, 25(4), 671–682. https://doi.org/10.1016/j.engappai.2010.06.002

Lu, Y. (2017b). Industry 4.0: A survey on technologies, applications and open research issues. Journal of Industrial Information Integration, 6, 1–10. https://doi.org/10.1016/j.jii.2017.04.005

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense passage retrieval for open-domain question answering. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 6769–6781. https://doi.org/10.48550/arXiv.2004.04906

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence. IEEE Access, 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Lu, Y., Liu, C., Kevin, I., Wang, K., Huang, H., & Xu, X. (2020). Digital twin-driven smart manufacturing: Connotation, reference model, applications and research issues. Robotics and Computer-Integrated Manufacturing, 61, 101837. https://doi.org/10.1016/j.rcim.2019.101837

Sankar, N. R., & Prabhu, B. S. (2001). Modified approach for prioritization of failures in a system failure mode and effects analysis. International Journal of Quality & Reliability Management, 18(3), 324–336. https://doi.org/10.1108/02656710110383737

Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., & Hajishirzi, H. (2023). When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 9802–9822. https://doi.org/10.48550/arXiv.2212.10511

Knowles, W., Prince, D., Hutchison, D., Disso, J. F. P., & Jones, K. (2015). A survey of cyber security management in industrial control systems. International Journal of Critical Infrastructure Protection, 9, 52–80. https://doi.org/10.1016/j.ijcip.2015.02.002

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), Article 93. https://doi.org/10.1145/3236009

Huang, Z., Shen, Y., Li, J., Fey, M., & Brecher, C. (2021). A survey on AI-driven digital twins in Industry 4.0: Smart manufacturing and advanced robotics. Sensors, 21(19), 6340. https://doi.org/10.3390/s21196340

Lu, Y. (2025). The current status and developing trends of Industry 4.0: A review. Information Systems Frontiers, 27(1), 215–234. https://doi.org/10.1007/s10796-021-10221-w

Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., van den Driessche, G. B., Lespiau, J.-B., Damoc, B., Clark, A., de Las Casas, D., Guy, A., Menick, J., Ring, R., Hennigan, T., Huang, S., Maggiore, L., Jones, C., Cassirer, A., ... Sifre, L. (2022). Improving language models by retrieving from trillions of tokens. Proceedings of the 39th International Conference on Machine Learning, 2206–2240. https://doi.org/10.48550/arXiv.2112.04426

Fenton, N. E., Neil, M., & Caballero, J. G. (2007). Using ranked nodes to model qualitative judgments in Bayesian networks. IEEE Transactions on Knowledge and Data Engineering, 19(10), 1420–1432. https://doi.org/10.1109/TKDE.2007.1073

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 30(3), 286–297. https://doi.org/10.1109/3468.844354

Lu, Y., & Zheng, X. (2020). 6G: A survey on technologies, scenarios, challenges, and the related issues. Journal of Industrial Information Integration, 19, 100158. https://doi.org/10.1016/j.jii.2020.100158

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterjee, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2021). On the opportunities and risks of foundation models. arXiv. https://doi.org/10.48550/arXiv.2108.07258

Mendling, J., Weber, I., van der Aalst, W., vom Brocke, J., Cabanillas, C., Daniel, F., Debois, S., Di Ciccio, C., Dumas, M., Dustdar, S., Gal, A., Garcia-Banuelos, L., Governatori, G., Hull, R., La Rosa, M., Leopold, H., Leymann, F., Recker, J., Reichert, M., Reijers, H. A., Rinderle-Ma, S., Solti, A., Rosemann, M., Schulte, S., Singh, M. P., Slaats, T., Staples, M., Weber, B., Weidlich, M., Weske, M., Xu, X., & Zhu, L. (2018). Blockchains for business process management: Challenges and opportunities. ACM Transactions on Management Information Systems, 9(1), Article 4. https://doi.org/10.1145/3183367

Giraldo, J., Urbina, D. I., Cárdenas, A. A., Valente, J., Faisman, B., Ruth, B., Tippenhauer, N. O., Sandberg, H., & Candell, R. (2018). A survey of physics-based attack detection in cyber-physical systems. ACM Computing Surveys, 51(4), Article 76. https://doi.org/10.1145/3203245

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

Lu, Y. (2021). Technological innovation and the emergence of a new interdisciplinary field: Management Analytics. Nanotechnologies in Construction, 13(3), 181–192. https://doi.org/10.15828/2075-8545-2021-13-3-181-192

Wan, J., Tang, S., Li, D., Wang, S., Liu, C., Abbas, H., & Vasilakos, A. V. (2017). A manufacturing big data solution for active preventive maintenance. IEEE Transactions on Industrial Informatics, 13(4), 2039–2047. https://doi.org/10.1109/TII.2017.2670505

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., ... Wen, J.-R. (2023). A survey of large language models. arXiv. https://doi.org/10.48550/arXiv.2303.18223

Ani, U. P. D., He, H., & Tiwari, A. (2017). Review of cybersecurity issues in industrial critical infrastructure: Manufacturing in perspective. Journal of Cyber Security Technology, 1(1), 32–74. https://doi.org/10.1080/23742917.2016.1252211

Jones, D., Snider, C., Nassehi, A., Yon, J., & Hicks, B. (2020). Characterising the digital twin: A systematic literature review. CIRP Journal of Manufacturing Science and Technology, 29, 36–52. https://doi.org/10.1016/j.cirpj.2020.02.002

Lu, Y., Sigov, A. S., Ratkin, L., Ivanov, L. A., & Zuo, M. (2023). Quantum computing and industrial information integration: A review. Journal of Industrial Information Integration, 35, 100511. https://doi.org/10.1016/j.jii.2023.100511

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–13. https://doi.org/10.1145/3290605.3300233

Aguilera, P. A., Fernández, A., Fernández, R., Rumí, R., & Salmerón, A. (2011). Bayesian networks in environmental modelling. Environmental Modelling & Software, 26(12), 1376–1388. https://doi.org/10.1016/j.envsoft.2011.06.004

Wu, H. P., Liu, Z., Dong, H. Y., Lu, Y., & Xu, L. D. (2025). Revolutionizing internal auditing: Harnessing the power of blockchain. Enterprise Information Systems, 19(1–2), 2448003. https://doi.org/10.1080/17517575.2024.2448003

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–14. https://doi.org/10.1145/3313831.3376219

Wang, F., & Lu, Y. (2026). Quantum information systems: Foundations, intelligent algorithms, and cross-industry applications. Journal of Emerging Technologies with Industrial Applications, 1(1), Article 3. https://doi.org/10.53941/jetia.2026.100003

Qi, Q., Tao, F., Hu, T., Anwer, N., Liu, A., Wei, Y., Wang, L., & Nee, A. Y. C. (2021). Enabling technologies and tools for digital twin. Journal of Manufacturing Systems, 58, 3–21. https://doi.org/10.1016/j.jmsy.2019.10.001

Lu, W., Lu, Y., Li, J., Sigov, A., Ratkin, L., & Ivanov, L. A. (2024a). Quantum machine learning: Classifications, challenges, and solutions. Journal of Industrial Information Integration, 42, 100736. https://doi.org/10.1016/j.jii.2024.100736

Ivanov, D. (2021). Digital supply chain management and technology to enhance resilience by building and using end-to-end visibility. IEEE Engineering Management Review, 49(1), 46–57. https://doi.org/10.1109/EMR.2021.3059871

Lu, Y., Pisarenko, Z. V., Yang, L., & Ye, C. (2024b). Advancing decision-making: The role of management analytics in modern business practices. Nanotechnologies in Construction, 16(5), 431–440. https://doi.org/10.15828/2075-8545-2024-16-5-431-440

Wang, J., Ma, Y., Zhang, L., Gao, R. X., & Wu, D. (2018). Deep learning for smart manufacturing: Methods and applications. Journal of Manufacturing Systems, 48, 144–156. https://doi.org/10.1016/j.jmsy.2018.01.003

Lu, Y. (2018). Blockchain and the related issues: A review of current research topics. Journal of Management Analytics, 5(4), 231–255. https://doi.org/10.1080/23270012.2018.1516523