

Evidence-Chain Trustworthiness Evaluation for Retrieval-Augmented Emergency Plan Generation in Hazardous Chemical Accidents

Yanglin Gao^{1*}

¹ School of Management, University of Shanghai for Science and Technology, Shanghai, China

*Correspondence: Yanglin Gao

Received: 4 June 2026; **Accepted:** 6 July 2026; **Published:** 16 July 2026

Abstract:

Emergency plan generation for hazardous chemical accidents is a high-risk decision-support task. A useful plan must be grounded in accident facts, hazardous-substance properties, operational constraints and regulatory requirements. Conventional retrieval-augmented generation (RAG) improves factual grounding, but retrieved fragments are often treated as independent context, which weakens control over evidence completeness, constraint consistency and traceability. This paper proposes an Evidence-Chain Trustworthiness (ECT) model for retrieval-augmented emergency plan generation. The model organizes domain evidence into four linked layers: accident scenario, hazardous substance, response measure and regulatory basis. It evaluates candidate evidence chains through five dimensions: semantic relevance, structural completeness, constraint consistency, source traceability and timeliness. The resulting score guides both plan generation and feedback-based re-retrieval when a draft lacks key constraints or legal support. A controlled simulation benchmark compares ECT-RAG with direct large language model generation, BM25-RAG, vector-RAG and graph-RAG. Results show that ECT-RAG achieves the highest overall trustworthiness, compliance consistency, traceability and robustness under incomplete accident descriptions. The study provides a practical route for safer and more auditable AI-assisted emergency management.

Keywords:

Hazardous Chemical Accident, Emergency Plan Generation, Evidence Chain, Knowledge Graph, Retrieval-Augmented Generation, Trustworthiness Evaluation

1. Introduction

Hazardous chemical accidents often develop suddenly and can combine leakage, fire, explosion, poisoning and environmental contamination. In this context, an emergency plan is not only a readable description of actions. It must explain why a response measure is applicable, which chemical constraints restrict the measure, and which rules or standards support the decision. A plan that overlooks chemical incompatibility or applies an unsuitable firefighting method may increase operational risk.

Large language models can organize accident information and produce fluent procedural text, but fluency alone is insufficient for emergency management. Unsupported recommendations, confused substance properties and weak regulatory grounding are unacceptable in high-risk response. RAG reduces hallucination by conditioning generation on external evidence [1][2]. However, basic sparse or dense retrieval normally ranks text chunks by keyword matching or semantic similarity. A chunk may be related to the accident while still missing the response boundary conditions that determine whether a measure is safe.

This paper focuses on evidence-chain trustworthiness. Instead of asking only whether retrieved evidence is relevant, the proposed model asks whether the evidence forms a complete, consistent and traceable decision-support chain. The main contribution is an ECT model that evaluates structured chains before generation and then uses draft-level review to trigger targeted re-retrieval. The approach is designed for hazardous chemical emergency planning, but the same idea can support other safety-critical RAG applications.

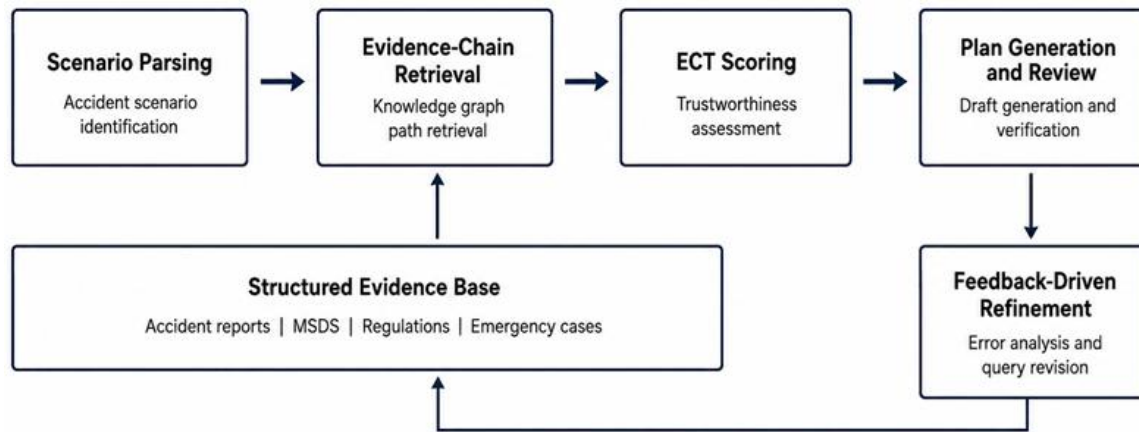
2. Related Work

RAG combines a parametric language model with an external knowledge source and has become a common method for knowledge-intensive generation. Dense passage retrieval and hybrid retrieval improve recall by representing queries and documents in semantic space [3]. More recent evaluation frameworks examine answer relevance, faithfulness and reasoning quality [4][5]. These approaches are useful, but domain-specific emergency planning also needs explicit checks for safety constraints, regulatory support and source traceability.

Knowledge graphs represent entities, attributes and relations in a structured form. In emergency management, they can connect accident types, hazardous substances, response resources, protective measures and standards. Graph-based RAG can retrieve paths rather than isolated passages, which makes decision logic more visible [6][7]. Existing emergency decision studies show the value of knowledge graphs for risk identification and aided decision making [8][9], but the generated plan still needs an evidence-level evaluation mechanism. A response measure is reliable only when it is linked to the relevant accident scenario, substance property and regulatory basis.

3. Evidence-Chain Trustworthiness Model

The proposed framework contains four modules: scenario parsing, evidence-chain retrieval, trustworthiness scoring and plan generation. The input is a natural-language accident description. Scenario parsing extracts the accident type, hazardous substance, location, exposure condition and missing elements. Evidence-chain retrieval searches a structured evidence base and constructs candidate chains linking scenario, substance, response measure and regulation. Trustworthiness scoring ranks and filters these chains before they are inserted into the prompt. Plan generation then produces a managerial emergency plan and activates correction if the draft fails evidence-based review. (Figure 1)



Output: An evidence-supported, regulation-aware, and traceable emergency response plan

Figure 1. Evidence-chain trustworthiness framework for emergency plan generation.

3.1. Structured Evidence Chains

An evidence chain is defined as a path that connects four knowledge layers. The scenario layer records the accident type and surrounding conditions. The substance layer records toxicity, flammability, reactivity, incompatibility and storage constraints. The measure layer records evacuation, isolation, leakage control, fire suppression, medical aid, disposal and command actions. The regulation layer records applicable standards, safety rules and emergency management requirements.

Let q denote the accident query and $C = \{c_1, c_2, \dots, c_n\}$ denote candidate evidence chains. A chain c_i consists of ordered nodes and relations that support a potential response decision. The model therefore avoids treating every retrieved fragment as equally useful. Evidence with high semantic similarity but low constraint consistency can be filtered out before it influences the generated plan.

Table 1. Evidence-chain elements and decision-support functions.

Layer	Main content	Decision-support function
Accident scenario	leakage, fire, explosion, poisoning, environmental release	defines response objective and urgency
Hazardous substance	toxicity, flammability, reactivity, incompatibility, storage condition	sets physical and chemical constraints
Response measure	evacuation, isolation, detection, fire control, disposal, medical aid	provides executable emergency actions
Regulatory basis	laws, standards, operating procedures and safety requirements	ensures compliance and traceability

3.2. Trustworthiness Scoring

The ECT score evaluates whether a candidate chain can support an emergency plan. Five dimensions are used: semantic relevance R , structural completeness L , constraint consistency K , source traceability T and timeliness U . Constraint consistency and traceability receive higher baseline weights because safety and compliance are decisive in hazardous chemical response. The score is calculated as: $ECT(c_i) = w_1R(c_i, q) + w_2L(c_i) + w_3K(c_i) + w_4T(c_i) + w_5U(c_i)$.

After draft generation, the system performs evidence-based review. If a plan mentions fire suppression but lacks a chemical-specific prohibition, the next retrieval

request emphasizes incompatibility and firefighting restrictions. If the plan lacks legal support, the revised query emphasizes standards and emergency provisions. This feedback loop repairs the evidence foundation rather than simply retrying generation with the same context.

Table 2. *Scoring dimensions of the ECT model.*

Dimension	Symbol	Evaluation focus	Weight
Semantic relevance	R	match between query and evidence chain	0.20
Structural completeness	L	coverage of scenario, substance, measure and regulation	0.20
Constraint consistency	K	consistency with chemical properties and safety prohibitions	0.25
Source traceability	T	clarity and credibility of evidence source	0.20
Timeliness	U	currency of standards, cases and records	0.15

4. Experimental Design

A controlled simulation benchmark was built around typical hazardous chemical accident scenarios, including leakage, fire, explosion, poisoning exposure and mixed-risk events. Each scenario contains an accident description, a hazardous substance, key response constraints and a reference evidence chain. To test robustness, four input-completeness levels were prepared: complete description, light missing information, moderate missing information and severe missing information.

Five methods were compared. Direct LLM generates a plan without retrieval. BM25-RAG retrieves sparse keyword matches [10]. Vector-RAG retrieves semantically similar evidence. Graph-RAG retrieves graph neighborhoods and relation paths. ECT-RAG combines graph retrieval with trustworthiness scoring and feedback-based re-retrieval. All methods used the same output template and were evaluated by the same rubric.

Evaluation includes evidence-level and plan-level indicators. Evidence-level indicators examine relevance, structural completeness and traceability. Plan-level indicators examine factual consistency, compliance consistency, effectiveness and robustness. Each generated plan was checked against the reference evidence chain, chemical constraints and regulatory requirements. The same random seed, temperature and maximum length were used across methods to reduce generation fluctuation.

Table 3. *Simulation benchmark setting.*

Item	Setting
Scenario categories	leakage, fire, explosion, poisoning exposure and mixed-risk events
Test queries	120 accident descriptions with controlled missing information
Evidence sources	accident reports, MSDS-style records, emergency cases and regulatory clauses
Baselines	Direct LLM, BM25-RAG, Vector-RAG and Graph-RAG
Indicators	trustworthiness, factual consistency, compliance, effectiveness and traceability

5. Results and Analysis

Retrieval improves emergency plan generation, but retrieval strategy matters. Direct LLM obtains the lowest scores because it lacks external evidence support. BM25-RAG improves factual grounding when the query contains clear chemical or

regulatory terms, but it is less stable when descriptions are incomplete. Vector-RAG improves semantic recall, yet it may return passages that are similar in wording but unsafe under specific chemical constraints. Graph-RAG performs better by using relations among scenario, substance and measure nodes. ECT-RAG achieves the best results because it further checks whether the retrieved evidence forms a complete and consistent decision-support chain.

Robustness is important because early accident reports are often incomplete. As information becomes less complete, all methods decline. The decline of ECT-RAG is smaller because the chain score preserves key substance constraints and regulatory support even when the initial query is weak. The advantage is especially clear under moderate and severe missing conditions, where keyword and vector retrieval are more likely to miss response boundary conditions.

The ablation analysis confirms the contribution of each component. Removing the knowledge graph causes the largest decline, showing that the four-layer structure is the foundation of reliable generation. Removing structural completeness weakens the connection between measures and regulations. Removing constraint consistency produces more unsafe or inappropriate actions. Removing source traceability reduces auditability. The improvement therefore comes from the combined effect of structured evidence and multi-dimensional evaluation.

Table 4. Performance comparison of generation methods.

Method	Trustworthiness	Factual consistency	Compliance consistency	Effectiveness	Traceability
Direct LLM	7.18	7.24	7.05	7.16	6.82
BM25-RAG	7.64	7.71	7.52	7.48	7.35
Vector-RAG	8.05	8.12	7.89	7.96	7.84
Graph-RAG	8.28	8.31	8.16	8.22	8.10
ECT-RAG	8.73	8.69	8.61	8.74	8.66

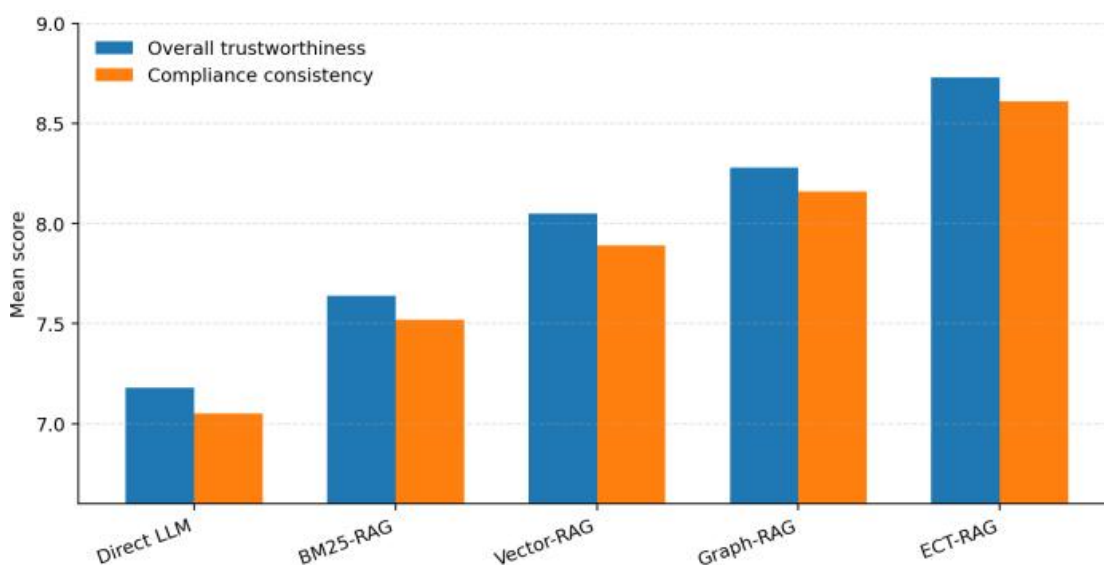


Figure 2. Comparison of overall trustworthiness and compliance consistency.

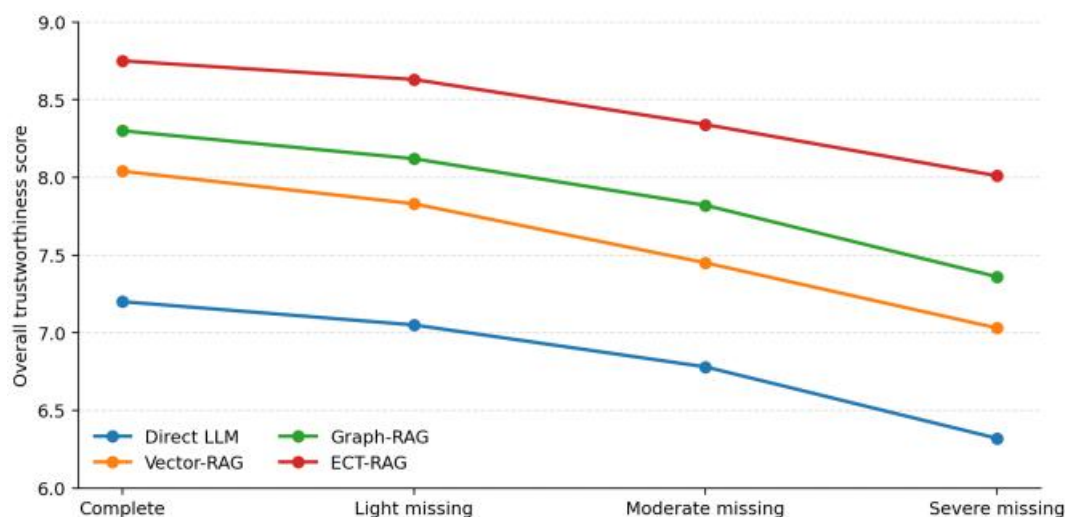


Figure 3. Robustness comparison under incomplete scenario descriptions.

Table 5. Ablation results of ECT-RAG.

Model variant	Trustworthiness	Compliance	Effectiveness	Traceability
ECT-RAG	8.73	8.61	8.74	8.66
w/o knowledge graph	7.86	7.72	7.78	7.45
w/o structural completeness	8.21	8.05	8.13	8.02
w/o constraint consistency	8.07	7.83	7.94	8.11
w/o source traceability	8.18	8.02	8.21	7.76

6. Discussion

The results suggest three implications for AI-assisted emergency management. First, evidence quality should be evaluated before generation. Semantically similar context may still be unsafe if it lacks chemical-specific restrictions. Second, regulatory evidence should be connected with response measures rather than appended as isolated references. A plan becomes easier to audit when each key action is supported by a visible chain. Third, feedback-based retrieval is valuable when accident descriptions are uncertain, because it converts generation failure into targeted evidence repair.

From a management perspective, the ECT model can support preparation of evidence-based response templates, review of AI-generated plans, and expert audit of safety-critical decisions. It does not replace professional judgment. Instead, it reduces the burden of searching fragmented knowledge and makes the evidence basis of generated measures more visible. The study is limited by the use of simulation scenarios. Real emergency response also involves sensor data, resource availability, weather, command coordination and legal responsibility. Future work should integrate real accident records, expert annotations and multi-source monitoring data.

7. Conclusions

This paper proposed an ECT evaluation model for retrieval-augmented emergency plan generation in hazardous chemical accidents. The model organizes accident scenario, hazardous substance, response measure and regulatory basis into structured

chains, and scores each chain through relevance, completeness, consistency, traceability and timeliness. The score guides generation and feedback-based re-retrieval.

The simulation experiment shows that ECT-RAG outperforms direct generation, BM25-RAG, vector-RAG and graph-RAG in trustworthiness, compliance consistency, traceability and robustness. Reliable emergency plan generation therefore depends not only on retrieving related evidence, but also on verifying whether the evidence can form a complete, consistent and traceable decision-support chain.

Conflicts of Interest

The author declares that there is no conflict of interest regarding the publication of this article.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

References

- [1] Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Kuttler, H.; Lewis, M.; Yih, W.T.; Rocktaschel, T.; Riedel, S.; Kiela, D. Retrieval-augmented generation for knowledge-intensive NLP tasks. *In Advances in Neural Information Processing Systems*, 2020, 33, 9459-9474.
- [2] Shuster, K.; Poff, S.; Chen, M.; Kiela, D.; Weston, J. Retrieval augmentation reduces hallucination in conversation. In M.-F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), *Findings of the Association for Computational Linguistics: EMNLP*, 2021; pp. 3784-3803. Association for Computational Linguistics. DOI: <https://doi.org/10.18653/v1/2021.findings-emnlp.320>.
- [3] Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; Yih, W.T. (2020). Dense passage retrieval for open-domain question answering. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 6769-6781). Association for Computational Linguistics. DOI: <https://doi.org/10.18653/v1/2020.emnlp-main.550>.
- [4] Es, S.; James, J.; Espinosa Anke, L.; Schockaert, S. (2024). RAGAs: Automated evaluation of retrieval augmented generation. In N. Aletras & O. De Clercq (Eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* (pp. 150-158). Association for Computational Linguistics. DOI: <https://doi.org/10.18653/v1/2024.eacl-demo.16>.
- [5] Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; Zhu, C. (2023). G-Eval: NLG evaluation using GPT-4 with better human alignment. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 2511-2522). Association for Computational Linguistics. DOI: <https://doi.org/10.18653/v1/2023.emnlp-main.153>.
- [6] Pan, S.; Luo, L.; Wang, Y.; Chen, C.; Wang, J.; Wu, X. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge*

- and Data Engineering*, 2024, 36(7), 3580-3599. DOI: <https://doi.org/10.1109/TKDE.2024.3352100>.
- [7] Edge, D.; Trinh, H.; Cheng, N.; Bradley, J.; Chao, A.; Mody, A.; Truitt, S.; Larson, J. (2024). From local to global: A Graph RAG approach to query-focused summarization. arXiv. <https://doi.org/10.48550/arXiv.2404.16130>.
- [8] Chen, M.; Tao, Z.; Tang, W.; Qin, T.; Yang, R.; Zhu, C. Enhancing emergency decision-making with knowledge graphs and large language models. *International Journal of Disaster Risk Reduction*, 2024, 113, Article 104804. DOI: <https://doi.org/10.1016/j.ijdr.2024.104804>.
- [9] Li, X.; Zhao, T.; Wen, J.; Cai, X. Many-objective emergency aided decision making based on knowledge graph. *Applied Intelligence*, 2024, 54, 7733-7749. DOI: <https://doi.org/10.1007/s10489-024-05557-0>.
- [10] Robertson, S.; Zaragoza, H. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 2009, 3(4), 333-389. DOI: <https://doi.org/10.1561/15000000019>.



© 2026 by the author(s); licensee International Technology and Science Publications (ITS), this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (<http://creativecommons.org/licenses/by/4.0/>)