

Crane Accident Investigation via LLM Reasoning: A System Thinking-guided Chain-of-Thought Approach

Junyu Chen¹ and Hung-Lin Chi¹

¹Department of Building and Real Estate, Faculty of Construction and Environment, The Hong Kong Polytechnic University, Hong Kong
junyu.chen@polyu.edu.hk, hung-lin.chi@polyu.edu.hk

Abstract –

Accident investigation is a critical component in construction safety management, yet conventional practices remain labor-intensive and time-consuming. This paper investigates the potential of Large Language Models (LLMs) to facilitate causal analysis of crane accidents by extracting and reasoning about contributing factors from narrative accident reports. Although LLMs have rapidly advanced, challenges persist in domain-specific reasoning, particularly in structured causal inference. To mitigate this limitation, we propose an HME-CoT prompting framework that integrates system thinking-based Human-Machine-Environment (HME) causation modeling with Chain of Thought (CoT) reasoning. This integration enables LLMs to systematically identify and classify causal factors relevant to crane accidents. Experiments conducted using DeepSeek-V3.2 on a curated accident dataset demonstrate that HME-CoT significantly enhances model reasoning performance, notably by improving the precision in identifying human errors and environmental constraints, compared with the basic zero-shot prompting strategy. These findings highlight the importance of embedding domain knowledge into prompt design and reaffirm the potential of LLM-based reasoning in supporting construction crane accident investigation.

Keywords –

Construction crane safety, Accident investigation, Large language models

1 Introduction

As one of the most important forms of lifting and transportation machinery, cranes play a pivotal role in construction operations [1]. Despite the indispensability, catastrophic crane-related accidents continue to occur with concerning frequency. Accident investigation serves as a fundamental mechanism for uncovering failure causes and informing future prevention strategies.

However, crane accidents typically arise from complex interactions among multifaceted factors, like human errors, mechanical problems, and environmental constraints. This multifactorial nature substantially complicates timely and accurate analysis. In practice, official investigations of severe crane accidents are often time-consuming and resource-intensive, frequently spanning several months. For instance, according to official press releases and preliminary investigation reports, the full investigation into the tower crane collapse at a construction site in Hong Kong on September 7, 2022, which resulted in three fatalities and six injuries, required approximately six months by the Labour Department [2] and up to twelve months by the Buildings Department [3]. Such extended investigation timelines delay the extraction of safety lessons and hinder the prompt implementation of effective preventive measures.

Deriving primary causal factors during the initial stage of accident investigation can provide valuable insights and effectively constrain the scope of subsequent in-depth inquiries. Given the inherent complexity of construction environments, system thinking has been widely recognized as an appropriate analytical foundation for accident investigation, as traditional linear causation models often fail to capture dynamic interactions across human, technical, and organizational dimensions [4]. Initial investigation reports, frequently disseminated through mass media channels, are typically based on self-reports and interviews with on-site personnel, including supervisors, engineers, and workers. Although these reports are less formal and structured than official investigation documents, they often contain rich contextual information. Such narratives can reveal both immediate hazards present at the time of accidents and latent decision-making processes of front-line practitioners, which are seldom fully captured through formal investigative procedures.

With recent advances in Natural Language Processing (NLP), the application of Large Language Models (LLMs) to accident analysis has gained increasing attention across safety-critical domains such

as aviation [5], transportation [6], and mining [7]. Compared with conventional rule-based or deep-learning NLP approaches, LLMs exhibit superior capability in handling free-text inputs and responding to open-ended queries, making them a promising tool for extracting salient information from unstructured initial accident reports. Nevertheless, accident investigation is not merely a text comprehension task; it requires domain-specific safety knowledge and structured, multi-step causal reasoning. Despite their strong language understanding capabilities, general-purpose LLMs often struggle with inferring complex causal chains underlying accidents. These limitations constrain the direct applicability of off-the-shelf LLMs to rigorous accident analysis without additional methodological guidance.

The Chain of Thought (CoT) prompting technique [8] has demonstrated effectiveness in enhancing LLM reasoning by decomposing complex tasks into explicit, structured intermediate steps. However, CoT strategies specifically tailored to crane accident investigation remain largely underexplored. Building on prior studies that incorporate domain knowledge into CoT prompting, this study embeds a crane accident causation model into the CoT reasoning process to facilitate systematic crane accident analysis. Man/Human-Machine-Environment (HME) system thinking [9] is adopted as the conceptual foundation, organizing accident causal factors into four hierarchical layers: crane-related human errors (H), mechanical problems (M), environmental constraints (E), and organizational management issues (O). Leveraging this structured representation, we develop an HME-guided CoT (HME-CoT) prompting framework that explicitly constrains and guides LLM reasoning. This framework enables domain-informed, multi-layer causal inference and supports the identification of primary causal factors during the initial stage of crane accident investigations.

The main contributions of this study are twofold: (1) integration of a crane accident causation model grounded in HME system thinking into the design of CoT prompts, thereby providing a structured approach to enhance the efficiency and rigor of crane accident investigations; and (2) empirical validation of the proposed HME-CoT framework using real-world crane accident cases from Hong Kong, demonstrating that LLMs can accurately identify multifaceted causal factors.

2 Literature Review

This section comprises two parts: First, studies on the causal analysis of crane accidents are reviewed, with primary emphasis on the development and application of crane accident causation models. Second, an overview of the use of LLMs in accident investigation is presented, focusing on techniques employed to enhance model

reasoning performance.

2.1 Causal Analysis of Crane Accidents

Research on crane safety has predominantly focused on identifying multifaceted causal factors that contribute to crane-related accidents. Prior studies indicate that human error accounts for approximately 43% of crane accidents, underscoring operator performance as a critical factor in crane safety [10]. For example, incidents involving workers being struck by moving machinery are frequently associated with operators' limited situational awareness [11]. In addition to human-related factors, mechanical problems (e.g., failures in control systems), environmental hazards (e.g., adverse weather conditions), and organizational issues (e.g., inadequate provision of personal protective equipment) have also been identified as significant contributors to crane accidents [12].

While identifying causal factors across multiple dimensions is essential, increasing attention has shifted toward understanding the causal chains and systematic interactions that underlie crane accidents. To this end, established accident causation frameworks, such as the Systems-Theoretic Accident Model and Process (STAMP) [13], the Human Factors Analysis and Classification System Model (HFACS) [14], and the Functional Resonance Analysis Method (FRAM) [15] have been applied to examine crane accident causation from a systems perspective. Among these approaches, HME system thinking, which emphasizes the holistic functioning of complex systems through interactions among human, mechanical, and environmental components, has been recognized as a particularly suitable framework for modeling crane accident causation [16]. Despite the valuable insights provided by existing studies, several limitations remain. First, many investigations focus on tracing causal factors associated with high-level system elements, such as government policies and regulatory enforcement, that are rarely emphasized during the initial investigation stage. Second, prior research has predominantly relied on structured data, thereby overlooking the rich information embedded in unstructured narratives such as initial investigation reports. These narratives often capture contextual cues and latent conditions that formal records may not fully reveal, yet are critical for early-stage accident analysis.

2.2 LLMs in Accident Investigation

LLMs, a class of Generative Artificial Intelligence (GAI), have shown great potential in accident analysis for hazard identification, accident classification, and causal inference [17]. CoT prompting [18], where intermediate reasoning steps are explicitly elicited, has been proposed as an effective prompting strategy for complex reasoning tasks. Compared with rationale-

augmented training and fine-tuning approaches, CoT prompting avoids the need for annotated rules and task-specific training data. Beyond prompting strategies, incorporating domain knowledge has been increasingly explored as a means of enhancing LLM performance in domain-specific reasoning tasks. In the context of accident analysis, prior studies have investigated the integration of structured causation models like HFACS to guide prompt design for extracting human factors from accident narratives [19]. Building on the previous works, the present study embeds system thinking-based crane accident causation modeling into the CoT prompting process, resulting in the proposed HME-CoT framework. This approach aims to enhance the reasoning capabilities of LLMs for systematic causal inference in crane accident investigations.

3 Proposed Method

3.1 HME-CoT Prompting Framework

Our proposed HME-CoT prompting framework is grounded in HME system thinking, which provides a structured and holistic approach for analyzing complex socio-technical systems. HME system thinking is a well-established paradigm in human factors and systems engineering, aimed at uncovering interactional patterns and latent risks that are often obscured when human-, machine-, and environment-related factors are examined in isolation. Building on this perspective, the causal factors of crane accidents are organized into four layers in a hierarchical structure. The machine dimension (M) captures explicit crane failure modes. The next layer represents human errors (H) that directly contribute to crane accidents. Subsequent layers contain preconditions leading to these human errors, including the constraints imposed by construction environments (E) and organizational management issues (O) involving broader tower crane stakeholders.

The prompts developed in this study focus on the first three layers, i.e., human, machine, and environment, by modeling how crane failure modes arise from human actions and environmental constraints. The organizational dimension remains in the conceptual HME hierarchy but is not operationalized here, because organizational and regulatory issues are often latent and cannot be inferred reliably from brief media-based initial reports. Given our focus on early-stage accident investigation from narrative reports, we prioritize dimensions that are directly observable from the available text. Within each dimension, sub-categories are defined based on a prior study on crane safety [12]. For instance, environmental constraints are decomposed into 1) poor weather conditions, 2) hazardous site conditions, 3) inadequate site information, 4) inadequate provision of

personal protective equipment, and 5) tight working schedules. Referencing these sub-categories, this study proposes the HME-CoT, which enables CoT reasoning aligned with HME system thinking. The HME-CoT prompt adopts a three-stage reasoning structure that guides LLMs to sequentially infer causal factors across HME dimensions from initial investigation reports of crane accidents.

Stage 1: Inferring Human Errors

In Stage 1, HME-CoT focuses on identifying human errors. Seven questions are generated to probe the narrative T for evidence of operator actions indicative of human errors, including poor communication, operator fatigue, carelessness, nonproficiency, poor vision, and poor on-site inspection or supervision. These questions are designed to capture deviations between expected and actual operations of crane operators within HME system thinking, while ensuring that each identified human error is explicitly grounded in accident investigation reports.

Stage 2: Inferring Crane Failure Modes

In Stage 2, HME-CoT identifies crane failure modes attributable to the operator actions inferred in Stage 1. Eight questions are designed to probe indications of mechanical problems, e.g., control system malfunction. These questions are formulated to capture discrepancies between the crane's expected and observed functional performance at the time of the accident, consistent with system-oriented accident analysis. All inferred failure modes are strictly grounded in factual descriptions and traceable from accident investigation reports.

Stage 3: Inferring Preconditions

Stage 3 aims to identify the preconditions that give rise to the human errors inferred in Stage 1. Drawing on HME system thinking, five questions explore causal factors from the environment dimension.

The collective question set Q , structured instructions I , accident narratives T , and the definitions of causal factors of crane accidents under HME-based categories C , are consolidated as input into the LLMs to generate the response Z , expressed as:

$$Z = LLM(Q, I, T, C) = \{(A_i, D_i, R_i) | i = 1, \dots, 20\} \quad (1)$$

where A_i indicates whether the answer to question Q_i is yes (denoted as '1') or no (denoted as '0'). When $A_i=1$, D_i identifies the most applicable HME factor code associated with Q_i , and R_i provides the corresponding justification for the selection.

Following the generation of Z , only the questions for which $A_i=1$, together with their associated codes D_i and explanations R_i are retained and compiled into the final analytical report:

$$Final\ Report = \{(Q_i, D_i, R_i) | A_i = 1\} \quad (2)$$

The complete question set Q and the instructions I

used in HME-CoT are presented in Table 1, showing the logical reasoning sequence underpinning the three-stage prompting process.

Table 1 Complete question set and structured instructions

Question	Content	Instruction
Human errors		
Q_1	Did the operator do or fail to do something that led to the accident occurring?	(1) As an expert in crane accident investigations, you are required to deduce the most applicable code for each of the following aspects: 1. Human errors by operators that contributed to the crane accident; 2. Crane failure modes attributable to the identified human errors; and 3. Environmental constraints that gave rise to the identified human errors, based on the following accident narrative: "{Text}." (2) Analyze in accordance with the classification of human errors, crane failure modes, and environmental constraints as provided: "{HME factor classification}." (3) Use the following steps to conduct a detailed accident analysis. Answer each question sequentially and select the most applicable code for each aspect. All analyses must be strictly grounded in statements from the accident narrative. (4) If the answer to Q_1 is "No," exit the process. (5) Proceed to the subsequent questions if at least one of Q_2 to Q_7 is answered "Yes."
Q_2	Did the error involve poor communication (H1) between operators and signallers?	
Q_3	Did the error involve operator fatigue (H2), which means continuous working over eight hours?	
Q_4	Did the error involve operator carelessness (H3), e.g., unintentional misoperation?	
Q_5	Did the error involve operator nonproficiency (H4), e.g., unprofessional conduct?	
Q_6	Did the error involve poor vision of operators (H5)?	
Q_7	Did the error involve poor inspection or supervision of operators (H6)?	
Failure modes		
Q_8	Did the accident involve workers being struck by cranes or crane parts (M1)?	(6) Start identifying the crane failure modes that are attributable to the identified human errors if any of the questions from Q_2 to Q_7 is answered "Yes."
Q_9	Did the accident involve workers being struck by objects from cranes (M2)?	
Q_{10}	Did the accident involve overload (M3)?	
Q_{11}	Did the accident involve malfunction of crane control systems (M4)?	
Q_{12}	Did the accident involve crane collapse (M5)?	
Q_{13}	Did the accident involve crane overturn (M6)?	
Q_{14}	Did the accident involve overhead power line electrocutions (M7)?	
Q_{15}	Did the accident involve operators or other workers being caught in or between (M8)?	
Preconditions		
Q_{16}	Did poor weather conditions influence the human actions (E1)?	(7) Start identifying the preconditions that give rise to the identified human errors if any of the questions from Q_2 to Q_7 is answered "Yes."
Q_{17}	Did hazardous site conditions influence the human actions (E2)?	
Q_{18}	Did inadequate site information influence the human actions (E3)?	
Q_{19}	Did inadequate provision of personal protective equipment influence the human actions (E4)?	

3.2 Data Collection and Processing

Due to the absence of publicly available datasets that jointly provide initial accident investigation reports and corresponding coded causal factors within the HME framework for crane accidents, this study constructed a dedicated dataset to support validation of the proposed HME-CoT prompting framework. Initial investigation reports of crane accidents occurring in Hong Kong between 2011 and 2020 were collected from publicly accessible sources, including reputable local news platforms (e.g., HK01). The raw dataset consists of unstructured accident narratives describing the accident sequence, operator actions, involved equipment, site conditions, and investigation findings. These narratives were first screened to remove records lacking sufficient details for causal analysis. Subsequently, an HME-based coding process was conducted by the authors to annotate human errors, crane failure modes, and environment-related preconditions. Records without identifiable human errors, mechanical problems, or environmental constraints were excluded from further analysis. The final dataset comprises 95 entries that link accident narratives with validated HME-based codes for causal factors, providing a reliable foundation for evaluating the performance of the proposed HME-CoT prompting framework.

3.3 Experiment Design

This study evaluated the effectiveness of the proposed HME-CoT prompting framework by comparing it against the basic zero-shot prompting strategy. The basic zero-shot prompt included only a brief task description, requiring the model to identify potential crane failure modes, human errors, and preconditions directly from the accident narrative. No examples or intermediate reasoning steps were provided; consequently, the model relied solely on its original language understanding and inference capabilities. The baseline was selected to represent commonly used prompt designs for information extraction and reasoning tasks, thereby providing a meaningful benchmark for assessing the added value of structured HME system thinking. In contrast, the proposed HME-CoT prompting framework explicitly embedded HME system thinking into the prompt design. By guiding the model through a predefined sequence of questions aligned with HME dimensions, HME-CoT constrained and structured the reasoning process, enabling systematic identification of causal factors. This design facilitated a direct and controlled comparison with unguided reasoning and

domain-informed CoT prompting grounded in HME system thinking.

3.4 Model Selection and Performance Metrics

As presented in Figure 1, an LLM-empowered web-based system, Agent Link for Crane Accident Investigation (AL4CAI), was developed using an integrated frontend-backend architecture. The frontend was implemented with Vue.js to provide a responsive and modular User Interface (UI), enabling safety investigators to input accident reports, select proper prompting strategies, and review model-generated analytical outputs. The backend was built using FastAPI in Python to handle request routing, data validation, and LLM inference workflows. The web-based application was deployed using Uvicorn, an ASGI server, to efficiently support concurrent requests and ensure real-time responsiveness. All experiments were conducted using DeepSeek-V3.2, accessed via its official API. As the latest LLM released by DeepSeek, the thinking mode of DeepSeek-V3.2, deepseek-reasoner, was selected for the accident investigation task due to its advanced reasoning capability and robustness in handling complex, multi-step prompting scenarios.

The performance of the prompting strategies was evaluated using standard classification metrics, including Accuracy, Precision, Recall, and F1-score. Each accident sample was assessed through a set of binary classification tasks corresponding to reasoning questions. Specifically, each sample comprised 19 classification tasks derived from 20 questions, with the first question serving as a screening item and therefore excluded from the performance evaluation. For each task, the model's binary response ("yes" or "no") was compared against the ground truth to determine its correctness. For each factor category, the metrics were computed as $\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$, $\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$, $\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$, and $\text{F1-score} = 2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall})$, where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. The average values reported for human errors (Ave_H), failure modes (Ave_M), and environmental constraints (Ave_E) were obtained by macro-averaging the category-wise scores within each group. In cases where a denominator became zero, the corresponding metric was treated as undefined and reported as "-" in

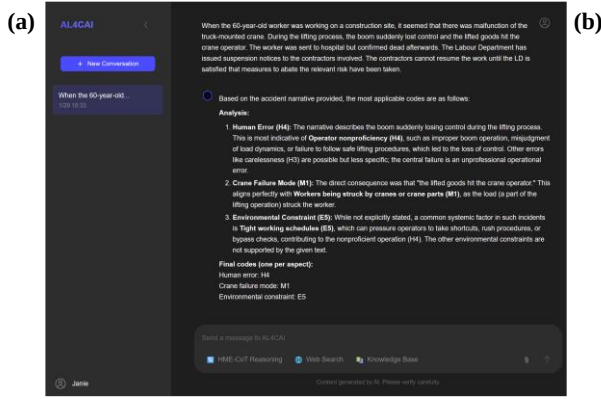


Figure 1. Accident narratives and responses using (a) basic zero-shot; (b) and (c) HME-CoT prompting

Table 2. For example, Precision is undefined when $TP + FP = 0$, and the F1-score is undefined when both Precision and Recall are zero or undefined. Undefined values were excluded from the relevant macro-average calculation. For illustration, if a factor category yields $TP=2$, $FP=1$, $FN=3$, and $TN=89$, then $Accuracy=0.96$, $Precision=0.67$, $Recall=0.40$, and $F1-score=0.50$.

4 Evaluation Results

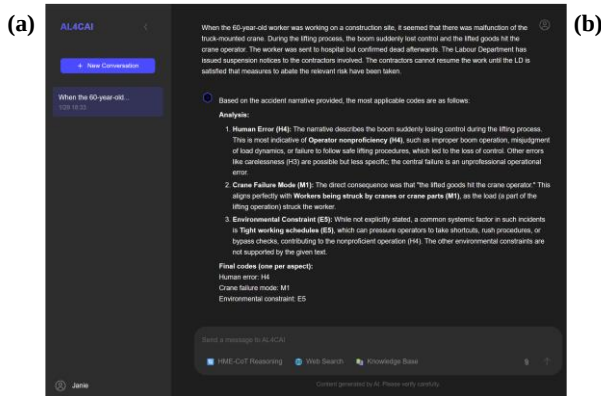


Figure 1. Accident narratives and responses using (a) basic zero-shot; (b) and (c) HME-CoT prompting

Table 2 reports a comparative evaluation of the basic zero-shot and HME-CoT prompting strategies on the dataset. For human error identification, the HME-CoT prompting strategy substantially outperformed the basic zero-shot approach, yielding a 32.7% improvement in average Precision. This indicates that HME-CoT is more effective at reducing false positives when detecting human errors. Both prompting strategies demonstrated strong performance in identifying communication-related issues between operators and signallers (H1), with Accuracy and Precision scores above 85%. In contrast, both approaches exhibited limited capability in

minimizing missed human errors, as reflected by average Recall scores below 50%. Notably, Recall dropped below 30% for both prompts in certain cases, highlighting a substantial number of undetected errors. Notably, the basic zero-shot prompt achieved a slightly higher average Recall for human error identification than HME-CoT (0.41 versus 0.39). This suggests a precision-recall trade-off: the HME-CoT prompt was better at suppressing unsupported human-error inferences, whereas the zero-shot prompt was comparatively more permissive and therefore retrieved slightly more potential positives at the cost of additional false positives. These results likely stem from the complex and implicit nature of human behavioral factors within narrative descriptions, which remain challenging for current LLMs to infer reliably.

For failure mode identification, the two prompting strategies exhibited comparable performance across all evaluation metrics, including Accuracy, Precision, Recall, and F1-score. While quantitative differences were marginal, HME-CoT offered enhanced interpretability by presenting more structured and transparent reasoning. Both strategies demonstrated strong performance in identifying struck-by objects (M2), overload (M3), overturn (M6), and electrocution events (M7), with the highest detection performance observed for M7. These failure modes were typically described in explicit and distinctive terms, which facilitated reliable recognition by LLMs. In contrast, both prompting strategies performed relatively poorly in detecting malfunction (M4) and collapse (M5). This performance degradation likely reflected the more ambiguous or indirect descriptions of these failure modes, which made them harder for the prompting strategies to infer correctly.

For environmental constraint identification, both prompting strategies performed well in identifying poor weather conditions (E1) and hazardous site conditions (E2). In contrast, their ability to detect environmental constraints related to inadequate site information (E3), inappropriate use of personal protective equipment (E4), and tight working schedules (E5) was relatively limited. In addition, HME-CoT outperformed the basic zero-shot prompting strategy in terms of Precision, achieving a 34.1% improvement in average Precision. Conversely, the basic zero-shot approach exhibited superior Recall, with a 31.8% improvement in average Recall over HME-CoT. These results indicate a trade-off between the two strategies: HME-CoT was more effective at reducing false positives through constrained, domain-informed reasoning, whereas the basic zero-shot approach tended to identify a broader set of potential environmental factors at the cost of reduced specificity.

5 Discussion and Conclusion

This study represents an initial effort to apply GAI to

the early-stage construction crane accident investigation. We explored the use of LLMs to infer human errors, crane failure modes, and environmental constraints from unstructured accident narratives. Guided by HME system thinking, domain knowledge was embedded into the CoT prompting design, resulting in the proposed HME-CoT framework. By structuring the reasoning process into sequential stages aligned with human, machine, and environment dimensions, HME-CoT enabled systematic and transparent causal inference. Experimental results demonstrated that integrating domain knowledge into prompt design substantially improved model performance, particularly by increasing precision in identifying human errors and environmental constraints.

Several limitations remain. First, the dataset was primarily derived from media-based reports, which may lack critical details available only in official investigation documents; as a result, the applicability of the proposed approach to comprehensive crane accident investigations remains underexplored. Second, although this study considered human errors, crane failure modes, and environmental constraints, organizational and regulatory factors were not explicitly modeled. These higher-level factors are often pivotal in systemic accident analysis but are difficult to infer reliably from short narrative reports alone. Third, while HME-CoT improved precision, recall for certain categories, particularly implicit human errors and ambiguous environmental constraints, remained relatively low, reflecting the inherent challenges faced by current LLMs in reasoning over indirect or incomplete descriptions. Fourth, because the 95 cases were collected from Hong Kong, the current evaluation does not yet establish whether the same prompt structure will generalize to other construction accident types, languages, or regulatory contexts. Future work will address these limitations by expanding both data sources and methodological scope.

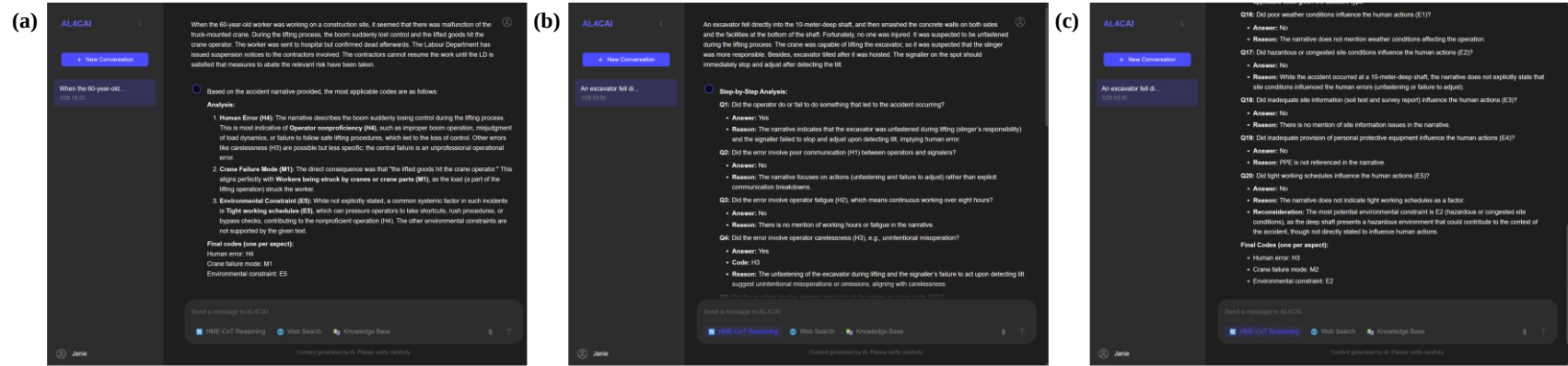


Figure 1. Accident narratives and responses using (a) basic zero-shot; (b) and (c) HME-CoT prompting

Table 2. Performance comparison of two prompting strategies

Matrix	Prompt	H1	H2	H3	H4	H5	H6	Ave_H	M1	M2	M3	M4
Accuracy	Basic zero-shot	0.86	0.94	0.62	0.82	0.95	0.74	0.82	0.80	0.87	0.96	0.95
	HME-CoT	0.87	0.95	0.57	0.78	0.97	0.77	0.82	0.83	0.83	0.96	0.93
Precision	Basic zero-shot	1.00	0.50	0.19	0.50	0.00	0.72	0.49	0.50	0.81	0.67	-
	HME-CoT	1.00	1.00	0.20	0.38	0.50	0.81	0.65	0.58	0.69	0.67	0.00
Recall	Basic zero-shot	0.24	0.67	0.50	0.47	0.00	0.55	0.41	0.68	0.81	0.67	0.00
	HME-CoT	0.29	0.17	0.64	0.35	0.33	0.55	0.39	0.58	0.87	0.67	0.00
F1-score	Basic zero-shot	0.38	0.57	0.28	0.48	-	0.63	0.47	0.58	0.81	0.67	-
	HME-CoT	0.45	0.29	0.31	0.36	0.40	0.66	0.41	0.58	0.77	0.67	-
Matrix	Prompt	M5	M6	M7	M8	Ave_M	E1	E2	E3	E4	E5	Ave_E
Accuracy	Basic zero-shot	0.94	0.93	1.00	0.93	0.92	0.96	0.69	0.95	0.96	0.77	0.87
	HME-CoT	0.94	0.92	1.00	0.93	0.92	0.98	0.69	0.96	0.98	0.74	0.87
Precision	Basic zero-shot	0.50	0.83	1.00	0.56	0.61	0.43	0.92	0.40	0.00	0.32	0.41
	HME-CoT	0.50	0.92	1.00	0.55	0.61	0.60	0.84	-	-	0.22	0.55
Recall	Basic zero-shot	0.33	0.79	1.00	0.63	0.61	1.00	0.66	0.50	0.00	0.75	0.58
	HME-CoT	0.33	0.63	1.00	0.75	0.60	1.00	0.76	0.00	0.00	0.42	0.44
F1-score	Basic zero-shot	0.40	0.81	1.00	0.59	0.61	0.60	0.77	0.44	-	0.45	0.57
	HME-CoT	0.40	0.75	1.00	0.63	0.60	0.75	0.79	-	-	0.29	0.61

6 Acknowledgements

The authors would like to thank the Research Grants Council, Hong Kong, for its funding support under the General Research Fund (PolyU 15223322).

References

- [1] A. Shapira, G. Lucko, and C.J. Schexnayder. Cranes for building construction projects, *Journal of Construction Engineering and Management*, 133(9):690–700, 2007.
- [2] Labour Department. Brief account of a fatal industrial accident involving the collapse of a tower crane at a construction site on Anderson Road on September 7, 2022. On-line: https://gia.info.gov.hk/general/202303/06/P2023030600356_414342_1_1678086026375.pdf, Accessed: 12/18/2025.
- [3] Buildings Department. Press releases. Buildings Department instigates prosecution regarding tower crane collapse incident on Anderson Road, Sau Mau Ping. On-line: <https://www.bd.gov.hk/en/whats-new/press-releases/2023/0906-buildings-department-instigates-prosecution-regarding-tower-crane-collapse-incident-on-anderson-road-sau-mau-ping.html>, Accessed: 12/18/2025.
- [4] N. Leveson. A new accident model for engineering safer systems. *Safety Science*, 42(4):237–270, 2004.
- [5] Q. Liu, F. Li, K.K.H. Ng, J. Han, and S. Feng. Accident investigation via LLMs reasoning: HFACS-guided Chain-of-Thoughts enhance general aviation safety. *Expert Systems With Applications*, 269(2025):126422.
- [6] H. Zhen and J.J. Yang. CrashSage: A large language model-centered framework for contextual and interpretable traffic crash analysis. *Artificial Intelligence for Transportation*, 3–4(2025):100030.
- [7] G. Du and A. Chen. Coal Mine Accident Risk Analysis with Large Language Models and Bayesian Networks. *Sustainability*, 17(5):1896, 2025.
- [8] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. Le, and D. Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 24824–24837, New Orleans, LA, USA, 2022.
- [9] C.-F. Chi, D. Sigmund, Y.-C. Lin, and C. G. Drury. The development of a scenario-based human-machine-environment-procedure (HMEP) classification scheme for the root cause analysis of helicopter accidents. *Applied Ergonomics*, 103(2022):103771.
- [10] J.E. Beavers, J.R. Moore, R. Rinehart, W.R. Schriver. Crane-related fatalities in the construction industry. *Journal of Construction Engineering and Management*, 132(9):901-910, 2006.
- [11] T. Cheng and J. Teizer. Modeling tower crane operator visibility to minimize the risk of limited situational awareness. *Journal of Computing in Civil Engineering*, 28(3):04014004, 2014.
- [12] J. Chen, H.-L. Chi, Q. Du, and P. Wu. Investigation of operational concerns of construction crane operators: An approach integrating factor clustering and prioritization. *Journal of Management in Engineering*, 38(4):04022020, 2022.
- [13] J. Yin, H.A. Aidi, A.H. Nuzul, and A.B. Nabliah. Development of a hoisting safety risk framework based on the STAMP theory and PLS-SEM method. *IEEE Access*, 12(2024):122998-123017.
- [14] J. Wang, F. Guo, Y. Song, Y. Liu, X. Hu, and C. Yuan. Safety risk assessment of prefabricated buildings hoisting construction: Based on IHFACS-ISAM-BN. *Buildings*, 12(6):811, 2022.
- [15] H. Zhong, L. Chen, M. F. Antwi-Afari, Z. Bao, and K. Chen. Dynamic risk assessment of tower crane operations by integrating functional resonance analysis method and Bayesian network. *Developments in the Built Environment*, 23(2025):100699.
- [16] W. Zhang, N. Xue, J. Zhang, and X. Zhang. Identification of critical causal factors and paths of tower-crane accidents in China through system thinking and complex networks. *Journal of construction engineering and management*, 147(12):04021174, 2021.
- [17] H. Zhen and J.J. Yang. CrashSage: A large language model-centered framework for contextual and interpretable traffic crash analysis, *Artificial Intelligence for Transportation*, 3-4 (2025):100030.
- [18] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E.H. Chi, Q.V. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. *NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems*, 1800(2022):24824-24837.
- [19] Q. Liu, F. Li, K.K.H. Ng, J. Han, and S. Feng. Accident investigation via LLMs reasoning: HFACS-guided Chain-of-Thoughts enhance general aviation safety. *Expert Systems with Applications*, 269(2025):126422.