Social & Occupational Ergonomics, Vol. 228, 2026, 59–68
https://doi.org/10.54941/ahfe1007936

Prototype of a Generative AI-Based Analogy Application for Human Error Case Analysis

Aoi Fujiwara[1], Yuka Banno[1], and Yusaku Okada[2]

[1]Graduate School of Science and Technology, Keio University, Japan
[2]Faculty of Science and Technology, Keio University, Japan

ABSTRACT

Research at the intersection of human factors analysis and large language models (LLMs) has grown rapidly in recent years; however, much of this work emphasizes automation and efficiency, evaluating success primarily through model-centric metrics. In contrast, this study reframes generative AI not as an automation tool for analysis but as a collaborative partner for cognitive stimulation, and proposes PromptWeave, a prompt-design methodology intended to expand, deepen, and transform an analyst's reasoning. We applied PromptWeave to industrial accident cases and conducted a quantitative evaluation using human-centered KPIs. The results indicate consistently high performance across all KPIs, supporting the utility of PromptWeave as a reproducible collaboration protocol executable on an LLM platform.

Keywords: Human error, Accident causes analysis, Generative AI

INTRODUCTION

For operating organizations, investigating accident causes and contributing factors is not merely a matter of producing recommendations to prevent recurrence; it is a practice that can and should be linked to frontline improvement, enhanced operational resilience, and the renewal and strengthening of organizational management. Achieving this requires analysis that does not collapse into individual blame, but instead spans procedures, equipment, the working environment, information flows, and organizational conditions, thereby requiring broad expertise in safety management, human factors, and management. Developing and retaining such analysts has historically required substantial time and financial investment.

With the rapid advances of LLMs, generative AI has become a prominent topic in accident/incident and human error analysis. A growing body of work aims to improve throughput and consistency by using LLMs to read narratives such as accident reports and near-miss descriptions, extract contributing factors, and classify them into established frameworks such as HFACS, with applications expanding across aviation, rail, construction, and UAV safety domains. While these studies make important contributions to reducing analytical burden, their evaluation typically centers on how accurately the AI can classify factors or the extent to which expert judgement can be

Received February 5, 2026; Revised April 6, 2026; Accepted April 26, 2026; Available online July 20, 2026

automated. Consequently, analysts are implicitly positioned as recipients of AI outputs, and the question of how generative AI reshapes their reasoning has remained underexamined.

Parallel trends can also be seen in Incident Response and Root Cause Analysis (RCA) for cloud and software failures, where research on automating cause identification and mitigation generation is advancing rapidly. These studies prioritize diagnostic accuracy and the reduction of response time (MTTR) as key metrics, aiming to provide immediate decision support in operational settings. However, even in this RCA domain, evaluations often center on whether the AI reached the "correct" cause, leaving the qualitative changes in the human inference process outside the scope of assessment.

Human error analysis is not merely an information-processing task; it is a cognitive activity in which the analyst reinterprets the situation, rationally understands the operator's behavior, and conceives countermeasures based on that understanding. In practice, it is far more critical for the analyst to acquire multifaceted perspectives, deeply comprehend the background context and cognitive burdens, and design actionable measures, rather than simply enumerating factors. Therefore, AI support systems must be designed with a focus on how to extend and deepen the analyst's own thinking, rather than merely delivering results. Yet, research that utilizes generative AI to expand the analyst's cognitive mode in human error cases—and evaluates its effectiveness using human-centered metrics—remains limited.

Against this background, this study positions generative AI not as an automation tool for accident analysis, but as a cognitive support partner intended to augment analysts' thinking. We propose PromptWeave, a prompt-design methodology that explicitly incorporates analogy as a core inferential resource within the analysis procedure to guide and support the development of human thought. Furthermore, we evaluate the effectiveness of PromptWeave based on changes in human analytical thinking rather than on the correctness of AI output. Unlike conventional LLM applications that rely on model-centric metrics such as classification accuracy, we introduce human-centered Key Performance Indicators (KPIs), such as the diversity of factors recalled by the analyst, the depth of causal explanations, and the quality of the countermeasures proposed.

The novelty of this study lies in its central inquiry: how human analysis can be transformed by generative AI, diverging from research that competes solely on AI performance improvements. From this perspective, we aim to present new design principles and evaluation frameworks to both the Human Factors and Human–AI Collaboration research communities.

METHODOLOGY

This study aims to develop a prompt structure (PromptWeave), executable on a general-purpose generative AI model (ChatGPT-5.2), that performs the basic functions of accident factor analysis and countermeasure planning based on HFACS and applied Human Factors research. Additionally, it implements information presentation intended to achieve the following three specific support functions:

1. Perspective Expansion Support: Supporting the expansion of the user's field of view by presenting multifaceted factors.
2. Imagination Elicitation Support: Stimulating imagination by describing the contextual background and mechanisms behind factors.
3. Countermeasure Ideation Support: Enhancing the user's ability to formulate countermeasures by prompting novel proposals across both hard (technical) and soft (operational) dimensions.

In the initial design phase, we provided the model with large volumes of context, including accident data and knowledge of countermeasures, but the output was biased toward direct factors, failing to sufficiently extract latent factors. Consequently, we explicitly defined viewpoints and inference procedures, iteratively revising the prompts while monitoring the outputs. This process suppressed generic and attention-capturing responses, eventually achieving a depth of factor description consistent with expert analysis. However, because these results still exhibited reproducibility issues dependent on the execution environment, we analyzed differences between the tuned model and the baseline in order to systematize the input design (prompts/dialogue protocols) required to bridge that gap.

The input design resulting from this approach does not aim to produce an immediate answer from a single instruction. Instead, it decomposes the analysis process into multiple phases, providing a reasoning scaffold that intervenes in the analyst's inference process to guide perspective expansion, imagination elicitation, and countermeasure ideation step by step (Table 1).

Table 1: Input design of PromptWeave.

Design Item	Content
Output Design Concept	• Phase A: Factor analysis focusing on the relationship between humans and tasks. • Phase B: Analysis of the structural impact of systems, environments, and rules. • Phase C: Redesign proposals for organizational operations, culture, and management structures. • *Each phase is designed not to be independent, but to capture the interaction between humans, equipment, procedures, environment, and organization.*
Cognitive Policy	• Avoid convergence on a single cause; emphasize grasping multi-layered and interactive structures. • Do not rely on superficial "human error terminology"; encourage analysis that includes background structures, cognitive load, and institutional factors. • Improvement proposals must combine technical measures (Hard) and operational measures (Soft).
Operational Considerations	• Avoid mimicking past reports or existing measures; encourage independent, reconstructive reasoning. • Emphasize describing the "structure that made the event possible" rather than the "chronology of the event." • Adopt a structure where the AI autonomously develops multi-faceted thinking.

(Continued)

Table 1: Continued.

Design Item	Content
Design Philosophy	• Presented as a design integrating factor extraction and countermeasure generation. • Supports the generation of emergent knowledge by expanding the user's scope of thought. • A novel approach treating the prompt itself as a "design object."

These design specifications were operationalized as constraints interpretable by generative AI, thereby implementing five control structures within the prompts (Table 2). By clarifying the objectives, constraints, and expected outputs for each phase, we formalized the method as a reusable analysis procedure.

Table 2: Control structures of PromptWeave.

Control Element	Content
5-Layer Analysis	1. Physical (Equipment, UI, Phenomena) 2. Information (Content, Granularity, Flow) 3. Temporal (Delay, Deterioration, Progression) 4. Psychological (Attentional Resources, Cognitive Fixation, Normalcy Bias, etc.) 5. Interaction (Mechanisms of combined elements)
4-Paragraph Structure	1. Origin (Why could this situation exist?) 2. Amplification (Which structures strengthened it?) 3. Manifestation (When and how did it surface?) 4. Event Connection (Where did it act in the current sequence?)
Prohibited Terms	*Carelessness / Assumption / Lack of experience / Operating error / Judgment error / Inadvertence / Degraded judgment / Lack of attention / Simple mistake / Lack of ability* Avoids thought-stopping via blame-attribution terms, forcing paraphrasing into background structures.
Factor Categories	1. L (Individual) 2. T (Team) 3. H (Hardware) 4. P (Procedures) 5. E (Environment) 6. C (Education) 7. R (Rules) 8. M (Management) 9. O (Org Design) 10. X (Other / Culture) Forces a multi-faceted perspective by using predefined multiple categories.

In practical use, prompts are interactively adapted to the industry domain and the operator's organizational scale.

EVALUATION

For the empirical evaluation of this study, we utilized the accident information database (PEC-SAFER) publicly available from the Japan Petroleum Energy Center (JPEC). The dataset comprised 121 industrial accident cases that occurred between 2006 and 2010.

Aligned with our objective to assess how the intervention of generative AI alters the analyst's thought process, we defined a set of human-centered Key Performance Indicators (KPIs). Unlike traditional metrics that measure the "correctness" of the output, these KPIs are designed to capture qualitative changes in the analysis process and its outcomes. Specifically, the indicators were constructed across four dimensions: (1) the diversity and comprehensiveness of factors recalled by the analyst, (2) the depth and consistency of causal explanations, (3) the specificity and scope of intervention levels in countermeasure planning, and (4) the degree of the analyst's subjective engagement.

1. **Perspective Expansion Support Metrics**
   - **Expert Factor Inclusion Rate:** The proportion of factors in the expert-defined set that were semantically covered by the output.
   - **Factor Bias Index:** The degree of distributional skew across the 10 factor categories (rated on a 5-point scale).
2. **Imagination Elicitation Support Metrics**
   - **Structural Depth Achievement:** The sufficiency of the four-layer analysis—Surface, Intermediate, Deep, and Event Connection (4 layers = 1.0, 3 layers = 0.75, 2 layers = 0.5, 1 layer = 0.25).
   - **Causal Reconstructability:** The extent to which a third party can rationally explain the process (causality) by which the operator's judgment was formed (5-point scale).
3. **Countermeasure Ideation Support Metrics**
   - **Behavior-Modifying Countermeasure Rate:** The percentage of measures that go beyond mere warnings to actively alter behaviors, judgments, or expectations.
   - **Timeline Design Maturity:** Whether the subject, role, and deadline (who/what/when) are explicitly defined for each measure (5-point scale).
4. **Human Resource Development & Cognitive Transformation Metrics**
   - **Analytical Text Improvement Index:** An assessment of the quality of the analysis text created using the output (rated on a 5-point scale based on causal depth, increased structural vocabulary, decreased personal blame expressions, and references to other categories).
   - **Output Intervention Rate:** The percentage of the AI output that was revised, added to, or rejected by the human analyst (target range: 30–60%).

This comparative design allows us to evaluate the impact of differences in analytical support structures on human analytical behavior, rather than differences in generative AI performance.

The evaluation framework of this study also considers the risk that the use of generative AI may overly steer analysts' thinking or lock them into a particular interpretation. To address this concern, some of the KPIs were designed to capture the extent of the analyst's active engagement, including the degree to which AI-generated outputs were revised, restructured, or otherwise reworked by the user. In addition, because the purpose of this study is to examine the effectiveness of the analytical support structure itself, the evaluation focuses on relative changes in analytical performance in order to minimize the influence of individual differences in participants' prior ability on the results.

For the evaluation experiment, two analytical conditions were prepared: a PromptWeave condition and a baseline condition for comparison. In the baseline condition, the same generative AI model and the same case descriptions as those used in the PromptWeave condition were employed, but no explicit control was imposed on the analytical procedure or reasoning structure. Instead, the model was prompted through a single-prompt instruction to conduct factor analysis and propose countermeasures for the case in question.

For all 121 target cases, analytical outputs were generated under both the baseline and PromptWeave conditions. These outputs were then assessed using the above evaluation metrics in accordance with the predefined scoring criteria. The evaluations were conducted independently by six safety practitioners drawn from three railway operators, two manufacturing companies, and one medical corporation, each of whom scored the full set of outputs. After confirming inter-rater agreement for each case using ICC(2, k) (two-way random-effects, absolute-agreement, average-measures), the mean score across the six raters was used as the representative value for each KPI in each case.

RESULTS AND DISCUSSION

Table 3 summarizes the evaluation results for the analytical outputs under each condition. For those metrics that reached the maximum score in all cases under the PromptWeave condition, only descriptive statistics for the two conditions are reported, as the results themselves indicate a consistently improved pattern. By contrast, for the two metrics that showed between-case variability under the PromptWeave condition—namely, the Behavior-Modifying Countermeasure Rate and the Output Intervention Rate—case-wise paired comparisons with the baseline condition were performed, and statistical significance was assessed using the Wilcoxon signed-rank test. Here, the Output Intervention Rate was treated not as a simple higher-is-better or lower-is-better measure, but as an optimal-range indicator, with 30–60% defined as the desirable range. Accordingly, for each case, the distance from this desirable range was calculated, and the paired differences between the baseline and PromptWeave conditions were tested using the Wilcoxon signed-rank test. The significance level was set at 1%, and Holm-adjusted p-values were used to control the family-wise error due to multiple testing (Table 4).

Table 3: Comparison of analytical output quality based on the evaluation metrics.

KPI	Baseline Median [Q1–Q3]	PromptWeave Median [Q1–Q3]
Expert Factor Inclusion Rate	0.404 [0.382–0.424]	100% in all cases
Factor Bias Index	3 [3.000–3.833]	5.0 (max) in all cases
Structural Depth Achievement	0.458 [0.292–0.500]	1.0 (max) in all cases
Causal Reconstructability	2 [1.833–2.167]	5.0 (max) in all cases
Behavior-Modifying Countermeasure Rate	0.317 [0.250–0.418]	0.840 [0.835–0.848]
Timeline Design Maturity	2 [2.000–2.833]	5.0 (max) in all cases
Analytical Text Improvement Index	1 [1.000–1.167]	5.0 (max) in all cases
Output Intervention Rate	0.680 [0.660–0.700]	0.427 [0.380–0.460]

Table 4: Results of the Wilcoxon signed-rank test with Holm correction.

KPI	Unadjusted *p*-value	Holm Correction Factor	Adjusted *p*-value
Behavior-Modifying Countermeasure Rate	1.325×10^{-21}	2	2.650×10^{-21}
Output Intervention Rate	1.700×10^{-21}	1	1.341×10^{-21}

To supplement the interpretation of the Behavior-Modifying Countermeasure Rate, the case-wise distributions of the proportions of countermeasure types A–D in the PromptWeave condition are presented as boxplots in a supplementary figure (Figure 1).

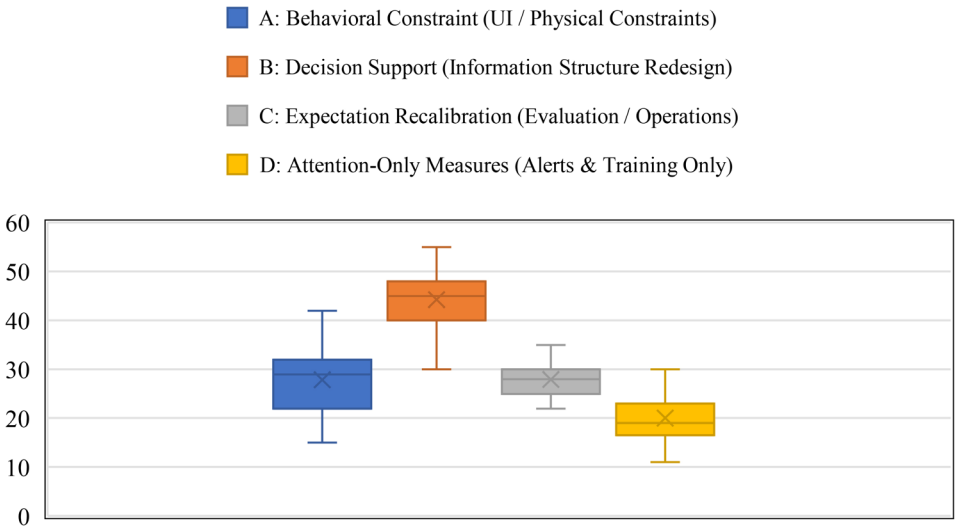

Figure 1: Distribution of generated countermeasures by category.

Compared with the condition using the baseline prompt, the finding that the PromptWeave condition consistently achieved high scores across all KPIs strongly suggests that merely introducing generative AI is not sufficient to realize such effects in analytical support. Rather, these results indicate that

the benefits are much more likely to emerge only when the support is designed to intervene in the analytical process itself. Of particular significance is that changes were consistently observed in the analyst's behavior and outcomes—dimensions that cannot be captured by traditional model-centric evaluation criteria such as "correctness" or "efficiency."

The fact that Causal Reconstructability and the Analytical Text Improvement Index reached high levels in all cases indicates that analysts were able to appropriate the presented analogies as part of their own reasoning. The analogies were internalized into the analysts' cognitive systems, enabling a multifaceted reconstruction of the rationality behind the operators' actions. This clearly demonstrates that the support structure effectively intervenes in the analytical process. This interactive process is expected to refine the analyst's mental model through repeated use.

Furthermore, the shift in countermeasure planning is critical from the perspective of practical value in human error analysis. In the baseline condition, there was a tendency to converge on stereotypical measures such as "pay more attention" or "re-education." In contrast, PromptWeave condition showed an increase in interventions addressing behavioral environments, information structures, and operational design. This suggests the potential for generative AI to fundamentally alter the quality of countermeasures.

It should be noted that the consistently high KPI scores observed in this study do not indicate the superiority of the underlying language model itself. Rather, they reflect the strong design constraints and structured reasoning scaffold imposed by PromptWeave, which intentionally shape the analyst's analytical process. In this sense, the results should be interpreted as evidence of the effectiveness of the collaboration protocol, not as a benchmark of generative AI performance.

At the same time, this study has several limitations. The evaluation was conducted with a relatively limited set of cases and participants, and the same effects cannot be assumed to generalize across all domains or levels of analytical expertise. In addition, because some of the KPI-based assessments involved the interpretation of written content, the ratings may have been influenced by evaluators' expectations or prior assumptions. Furthermore, for several KPIs, scores in the PromptWeave condition clustered near the upper bound, raising the possibility of ceiling effects that made it more difficult to capture between-condition differences and the magnitude of improvement with sufficient sensitivity. These issues suggest that further refinement and elaboration of the KPI design will be necessary. Moreover, precisely because the PromptWeave structure exerts a strong guiding effect, the risk of cognitive lock-in cannot be theoretically ruled out—that is, the possibility that analysts' thinking may converge too narrowly on particular perspectives or interpretive frames, leading them to overlook alternative lines of understanding. These limitations, however, should not be seen as undermining the value of the study. Rather, they should be understood as clarifying the design responsibilities that arise when generative AI is used as a tool for cognitive support.

CONCLUSION

This study proposed and developed PromptWeave, a protocol design methodology for treating generative AI not as a mere automation tool for human error case analysis, but as a collaborative partner that supports transformations in users' cognition, reasoning, and behavior. An implementation-based evaluation using industrial accident cases demonstrated that PromptWeave consistently promoted the expansion of the analyst's perspective, the deepening of causal understanding, and the widening of the countermeasure ideation space.

This research articulates a human-centered stance on designing collaborative structures centered on human thinking, demonstrating the potential of generative AI as an apparatus that structures and activates human cognitive activity. PromptWeave is not merely an analysis support tool; rather, it provides a methodological framework for designing and evaluating the quality of Human–AI collaboration.

Future work will aim to build a sustainable system that contributes to organization-wide human resource development and the fostering of safety culture by implementing output adjustments tailored to user attributes such as analytical experience and job role. A broader examination of PromptWeave's effectiveness and limitations is required through application to analysts with varying domains and proficiency levels, as well as through its deployment in education and training contexts.

REFERENCES

Ahmadi, A. et al. (2025) Improving Aviation Safety Analysis: Automated HFACS Classification Using Reinforcement Learning with Group Relative Policy Optimization.

Ahmadi, E. et al. (2025) Automatic Construction Accident Report Analysis Using Large Language Models (LLMs). Journal of Intelligent Construction. 3 (1), 1–10.

Ahmed, T. et al. (2023) 'Recommending Root-Cause and Mitigation Steps for Cloud Incidents Using Large Language Models', in Proceedings / International Conference on Software Engineering. 2023 Piscataway, NJ, USA: IEEE Press. pp. 1737–1749.

Chen, Y. et al. (2024) 'Automatic Root Cause Analysis via Large Language Models for Cloud Incidents', in EuroSys 2024 - Proceedings of the 2024 European Conference on Computer Systems. 2024 New York, NY, USA: ACM. pp. 674–688.

Fang, A. et al. (2025) A Goal-Driven Survey on Root Cause Analysis.

Li, Y. et al. (2025) 'COCA: Generative Root Cause Analysis for Distributed Systems with Code Knowledge', in Proceedings / International Conference on Software Engineering. 2025 Piscataway, NJ, USA: IEEE Press. pp. 1346–1358.

Liu, Q. et al. (2025) Accident investigation via LLMs reasoning: HFACS-guided Chain-of-Thoughts enhance general aviation safety. Expert systems with applications. 269.

Strong, J. et al. (2024) Trustworthy and Practical AI for Healthcare: A Guided Deferral System with Large Language Models.

Vats, V. et al. (2025) A Survey on Human-AI Collaboration with Large Foundation Models.

Wang, Z. et al. (2024) 'RCAgent: Cloud Root Cause Analysis by Autonomous Agents with Tool-Augmented Large Language Models', in PROCEEDINGS OF THE 33RD ACM INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, CIKM 2024. 2024 New York, NY, USA: ACM. pp. 4966–4974.

Xu, G. et al. (2025) Enhancing Intuitive Decision-Making and Reliance Through Human–AI Collaboration: A Review. Informatics (Basel). 12 (4), .

Yan, Y. et al. (2025) UAV Accident Forensics via HFACS-LLM Reasoning: Low-Altitude Safety Insights. Drones (Basel). 9 (10).