Human Systems Engineering and Design (IHSED 2026), Vol. 237*,* 2026, 619–626
https://doi.org/10.54941/ahfe1008201

The Effectiveness of Safety Observation Classification: A Case Study

Johanna Pulkkinen, Maria Lindholm, and Noora Nenonen

Tampere University, Tampere, Finland

ABSTRACT

Safety observations are an essential means of improving safety management and culture, and they play a key role in supporting the execution of occupational safety strategies. Information from observations can be obtained through, for example, case descriptions and various types of classifications. Classification enables the observations to be examined in a more structured manner. In this study, the applicability and effectiveness of the observation type classifications selected by the person who submitted the observation were analysed. The findings are based on an analysis of 1,988 safety observations made by a scaffolding service company. The observations were analysed to determine what proportion of the items classified into each category had been classified correctly, and what proportion of all observations belonging to a given category could be retrieved by examining that category. In addition, a generative artificial intelligence (GAI) was used to test how well it can categorise safety observations (n = 100). Based on the results, the individuals submitting the observations were generally able to classify them correctly when the category was easy to understand. Some categories were regarded as the most critical for the company's operations. The safety observation classification performed by the GAI was slightly weaker compared to the original human-made classifications. The findings also indicated that, in order to make better use of the information obtained through classification, general classifications should be adapted to company- or industry-specific categories. This will ensure that the information obtained is more specific and relevant.

Keywords: Generative artificial intelligence, Near miss, Occupational health and safety, Safety management

INTRODUCTION

Safety observations constitute an important element of proactive safety management, aiming to improve safety performance and prevent accidents before they occur. The practice of collecting safety observations has become more widespread across workplaces in different sectors (Gnoni et al., 2022), although substantial differences exist in how these practices are implemented across organisations (Gnoni and Saleh, 2017).

Safety observations typically capture unsafe conditions and unsafe acts, whereas near-miss incidents refer to unplanned events that had the potential to cause injury or damage to people, property, equipment, materials, or the environment but ultimately did not result in harm (Gnoni and Lettera, 2012; Wincek, 2016). Both safety observation and near-miss reporting systems can contribute valuable information for accident prevention and safety

Received April 21, 2026 Revised June 24, 2026 Accepted July 10, 2026 Available online September 25, 2026

management, support improvements in organisational safety performance, maintain safety awareness, and promote employee involvement in safety-related activities (Cambraia et al., 2010; Gnoni et al., 2022; Gnoni and Saleh, 2017). While safety observation systems can yield positive effects, poorly designed or implemented practices may also lead to unintended negative consequences, such as undermining trust between employees and management (Oswald et al., 2018).

The classification of incidents (including near misses) is necessary to enable appropriate follow-up actions and facilitate the identification of trends (Morrison, 2004). The categorisation of safety observations and near-miss incidents is not standardised, as both the number of categories and their definitions differ substantially across organisations and industrial sectors (Bicen and Celik, 2022; Fabiano and Currò, 2012; Fang et al., 2020). Safety observations can be classified in several ways. They are commonly categorised, for example, based on the nature of the observation, the consequences, the type of hazard, the severity, and the work phase or task involved. Event time and location are also frequently used as classification criteria (Al Shaaili et al., 2023; Gnoni and Lettera, 2012). Human factors analysis and classification system levels (Bicen and Celik, 2022), the EU Major Accident Reporting System approach (Fabiano and Currò, 2012), and the National Examination Board in Occupational Safety and Health categories (Al Shaaili et al., 2023) have all been utilised for near-miss categorisation, classification, and analysis.

For organisations that manage large safety databases, accurately classifying accident and near-miss narratives is a time-consuming task (Goh and Ubeynarayana, 2017). Variations may occur in the human-based classification of near-miss incidents, which can result in similar events being classified in different ways (Fang et al., 2020). In addition, the analysis of near-miss reports can be resource-intensive, which has led to growing interest in the use of information technology (IT) and artificial intelligence (AI) to support this process. However, current technological solutions still pose challenges, as they can result in misclassifications and raise concerns regarding the reliability of reported data (Goh and Ubeynarayana, 2017; Woźniak and Hoła, 2024).

This study aimed to analyse the applicability and effectiveness of the observation type classifications selected by the person who submitted the observation. A further aim was to compare the classifications produced by AI with those made by human assessors. The AI–based tests were conducted using a small sample of observations.

MATERIALS AND METHODS

The findings of this study are based on an analysis of safety observations made by a scaffolding service company. The case organisation operates in the industrial scaffolding sector and employs approximately 250 in-house workers, as well as approximately 400 subcontracted workers each year. The working environment is highly multilingual, with roughly 10 different languages in daily use. Annually, about 2,000 safety observations are reported, and this number has increased markedly in recent years. Each submitted observation includes detailed information such as the location, the type of observation, contributing factors, the employer involved, optional information about the observer, and a description of the surrounding circumstances. The case organisation uses a general classification scheme across all organisations. For each observation, the reporter selects the type of observation from a list of nine predefined categories.

The safety observation data used in this study cover a 14-month timeframe (March 2024 to April 2025). Positive observations and duplicates were excluded from the original dataset (N = 2,100). In total, 1,988 observations were analysed to determine what proportion of the items classified into each category had been correctly classified and what proportion of all observations belonging to a given category could be retrieved by examining that category. Before the analysis was undertaken, the classifications of the observations made by the reporting individuals were manually checked and corrected. The analysis results were discussed with company representatives in joint meetings, which helped enhance shared understanding and ensured that the company's perspectives were taken into account. Based on the analysis and the company's own assessments, the case company revised its existing categories to better reflect the needs of both the industry and the company.

In addition, 100 safety observations from April 2024 were loaded into Microsoft 365 Copilot Chat, which is based on the GPT-5 chat model. This generative artificial intelligence (GAI) was used to test how well it could categorise safety observations. The analysis was conducted in March 2026. The GAI was provided with the categories used in the case company and then instructed to categorise the 100 safety observations. The prompt discussion is presented in Table 1. All the safety observations were written in Finnish, and the prompts were also given in Finnish.

Table 1: Prompt discussion.

Prompt	GAI's Answer
The company uses the following factors to classify what has caused the situation described in the safety observation.	Here is the next step: I can now classify each observation according to the classification you gave, but I still need one clarification so that the result will surely correspond to the logic that is used in your company.
Classification: Personal protective equipment, clothing; Orderliness, cleanliness; Chemicals, air purity; Machinery, equipment, scaffolding, tools; Transport, warehouse work, other logistics work; Passageways, exit routes; Noise, lighting, temperature; Risk-taking; Weather conditions	I need one clarification from you before I do the classification: Do you want that: A) Each observation gets only one main class? – I choose the one that is most clearly the main cause of the situation. (E.g. slips → "Weather conditions") or

(Continued)

Table 1: Continued.

Prompt	GAI's Answer
Now examine the safety observations in the file. Read from each row the description of the observation (Column B) and classify each observation according to the classification I listed above.	B) Each observation can get more than one class? – If the observation relates for example both to walkways and to risk-taking, both are marked. When I know whether one or more than one class per observation is used, I will do the classification for the whole file.
	Which option (A or B)?
Use option A*	All right – I will use option A, i.e. for each observation only one main class is selected that best describes the cause of the situation. I can now do the classification automatically, but before I start, I still need one piece of information from you: Do you want that: 1) I do the classification and write the results here in the chat message in list form? – For each row: "Observation X → Class Y" or 2) You want me to add the classification back into Excel as a new column? – And I return to you the finished Excel file. Choose 1 or 2, and I will do the classification immediately.
Add the classification to Excel as a new column	Done! I have added the classification as a new column to the Excel file and formed a finished file.

*Option A was chosen because the case company used only one class for each observation.

RESULTS

Applicability and Effectiveness of the Observation Type Classifications

The classification performance varied across the categories. Table 2 summarises the share of observations across categories, the recall (the proportion of observations correctly classified within each category), and the precision (the proportion of observations assigned to a category that truly belong to that category).

Table 2: Recall and precision of the classification results.

Categories	Share	Recall	Precision
Personal protective equipment, clothing	5.4%	93.4%	77.3%
Orderliness, cleanliness	21.0%	91.7%	91.5%
Chemicals, air purity	1.3%	84.6%	64.7%

(Continued)

Table 2: Continued.

Categories	Share	Recall	Precision
Machinery, equipment, scaffolding, tools	29.0%	83.0%	97.3%
Transport, warehouse work, other logistics work	5.5%	93.5%	83.5%
Passageways, exit routes	12.9%	94.1%	98.3%
Noise, lighting, temperature	1.7%	72.7%	64.9%
Risk-taking	18.3%	95.6%	98.3%
Weather conditions	3.5%	62.3%	87.8%

For the study period, the largest share of observations was associated with "Machinery, equipment, scaffolding, tools" (29.0%), followed by "Orderliness, cleanliness" (21.0%) and "Risk-taking" (18.3%). The lowest proportions of observations were related to "Chemicals, air purity" (1.3%), "Noise, lighting, temperature" (1.7%), and "Weather conditions" (3.5%).

Overall, recall was high for most categories, particularly for "Risk-taking" (95.6%), "Passageways, exit routes" (94.1%), and "Personal protective equipment" (93.4%). Recall for observations related to "Weather conditions" (62.3%) and "Noise, lighting, temperature" (72.7%) were comparatively lower than for other categories.

Precision was especially high in categories related to "Machinery, equipment, scaffolding, tools" (97.3%), "Risk-taking" (98.3%), and "Passageways, exit routes" (98.3%). In contrast, precision was low in the categories "Chemicals, air purity" (64.7%) and "Noise, lighting, temperature" (64.9%).

Based on the analyses, the original classifications were revised to better reflect the needs of the construction industry and the company. The revised categories are: Working conditions; Orderliness and tidiness of working environment; Personal protective equipment and clothing; Tools and equipment; Scaffolding and weather protection structures; Vehicles, traffic, and logistics; Packaging and storage; Review of work practices; and Other.

Classifications Produced by Artificial Intelligence Compared With Those Made by Human Assessors

In 21% of cases, the GAI classified the safety observation in the same way as a human. Based on the researchers' assessment, the GAI's classification was judged to be better in a further 21% of cases. For example, for the observation "The sanding of the yard is inadequate. Several times at the same worksite I have almost slipped", the GAI classified it as "Weather conditions", while the original classification was "Orderliness, cleanliness".

The original, user-made classification was considered better in 33% of cases. For example, for the observation "Anchoring inspected and found to be in accordance with the plan", the original classification was "Machines, equipment, scaffolds, tools", while the GAI classified it as "Other".

In 20% of cases, either classification could have been chosen. For example, “An end flange was left on the stairs” was classified as “Passageways, exit routes” by GAI and as “Orderliness, cleanliness” by the person who wrote the observation.

Lastly, in 5% of cases, the researcher did not agree with either classification. For example, the observation “On the mill side of the project, at the staircase access and entrance to Company X’s office facilities, there is a tripping hazard due to an excessively large gap between the step and the site cabin” was classified as “Weather conditions” by GAI and as “Machinery, equipment, scaffolding, tools” in the original classification. However, the researchers would have classified this as “Passageways, exit routes”.

DISCUSSION AND CONCLUSION

Previous research has shown that the categorisation of safety observations is not standardised across organisations or industries and that human-based classification may be both time-consuming and subject to variation, particularly when dealing with large volumes of narrative data (Bicen and Celik, 2022; Fabiano and Currò, 2012; Goh and Ubeynarayana, 2017). These challenges underscore the inherent complexity of safety observation classification and the potential for inconsistency when category definitions are unclear or overlapping (Fang et al., 2020).

The findings of this study indicate that employees were generally able to classify safety observations correctly when the categories were clearly defined and aligned with familiar, industry-relevant hazards. Categories related to machinery, equipment, scaffolding, and tools were the most frequently used, reflecting the operational context of the case company and the relative ease of identifying traditional risk types. In contrast, categories associated with chemicals, air purity and noise, lightning, and temperature were less frequently used and showed lower recall and precision, suggesting difficulties in consistently interpreting and distinguishing these types of hazards.

The results further suggest that categories with lower classification accuracy may be insufficiently defined or may contain observations that could reasonably belong to multiple categories (Fang et al., 2020). To address these challenges, the company refined certain industry-specific categories, and an “Other” category was included to reduce forced or inaccurate classifications. Overall, the findings highlight that general classification schemes are most effective when adapted to company- or industry-specific contexts (Bicen and Celik, 2022; Fabiano and Currò, 2012). Such customisation improves the relevance and usability of classified safety data and supports more meaningful analysis and targeted safety improvements.

The findings also indicate that generative artificial intelligence (GAI) can classify safety observations at a level largely comparable to human assessors. In over 40% of the cases analysed in this study, the GAI’s classification was either identical to or more appropriate than the original human-made classification, demonstrating its potential as a supportive tool in safety data analysis. Nevertheless, consistent with earlier research, current technological solutions still present challenges, as misclassifications may occur and concerns remain regarding the reliability of reported data (Goh and Ubeynarayana, 2017; Woźniak and Hoła, 2024). The fact that human classifications were judged superior in approximately one-third of the cases underscores the continued importance of contextual understanding and expert judgement. Consequently, AI-based classification appears most suitable as part of a combined approach in which automated analysis is complemented by human oversight. In the present study, AI did not understand the terminology in all the safety observations, which decreased its accuracy. With a reasonable level of AI training, classification performance could be further improved, enabling the system to better learn and handle more challenging categories.

The findings of this study should be interpreted in light of certain limitations. The analysis was based on a single case company, which limits the generalisability of the results and highlights the need for future studies across multiple organisations and industries. In addition, the AI-based classification was conducted using a single GAI tool (Microsoft Copilot), without task-specific training and on a relatively small dataset. The results may not fully reflect the potential of GAI approaches when applied to larger datasets or with comprehensive AI model training.

To conclude, this study described how individuals making safety observations are generally able to classify observations correctly if the category is easy to understand. Furthermore, the study provided an example of how to test the functionality of the categorisation. Some categories were regarded as the most critical for the company's operations. Based on the results, some modifications were made to the categories. Lastly, testing was conducted to establish how well GAI performs in classifying safety observations. The results indicated that the GAI performed slightly weaker compared to the human-made original classifications. However, with more rigorous training of AI, these results may change.

ACKNOWLEDGMENT

The authors acknowledge the case company for providing the safety observation data. Additionally, the authors are grateful to Heini Ihalainen for analysing the case company's safety observations.

REFERENCES

Al Shaaili, M., Al Alawi, M., Ekyalimpa, R., Al Mawli, B., Al-Mamun, A. & Al Shahri, M. (2023). Near-miss accidents data analysis and knowledge dissemination in water construction projects in Oman, Heliyon, Volume 9, No. 11, e21607. https://doi.org/10.1016/j.heliyon.2023.e21607

Bicen, S. & Celik, M. (2022). A hybrid approach to near-miss report investigation towards next-generation safety solutions on-board ships, Ocean Engineering, Volume 266.

Cambraia, F.B., Saurin, T.A., & Formoso, C.T. (2010). Identification, analysis and dissemination of information on near misses: A case study in the construction industry, Safety Science, Volume 48, 91–99.

Fabiano, B. & Currò, F. (2012). From a survey on accidents in the downstream oil industry to the development of a detailed near-miss reporting system, Process Safety and Environmental Protection, Volume 90, No. 5, 357–367.

Fang, W., Luo, H., Xu, S., Love, P.E.D., Lu, Z., and Ye, C. (2020). Automated text classification of near-misses from safety reports: An improved deep learning approach, Advanced Engineering Informatic, Volume 44.

Gnoni, M.G. and Lettera, G. (2012). Near-miss management systems: A methodological comparison, Journal of Loss Prevention in the Process Industries, Volume 25, No. 3, 609–616. https://doi.org/10.1016/j.jlp.2012.01.005

Gnoni, M.G. and Saleh, J.H., (2017). Near miss management systems and observability-in-depth: Handling safety incidents and accident precursors in light of safety principles, Safety Science, Volume 91, 154–167. https://doi.org/10.1016/j.ssci.2016.08.012

Gnoni, M.G., Tornese, F., Guglielmi, A., Pellicci, M., Campo, G., and De Merich, D. (2022). Near miss management systems in the industrial sector: A literature review, Safety Science, Volume 150, 105704. https://doi.org/10.1016/j.ssci.2022.105704

Goh, Y.M. and Ubeynarayana, C.U. (2017). Construction accident narrative classification: An evaluation of text mining techniques, Accident Analysis and Prevention, Volume 108, 122–130. https://doi.org/10.1016/j.aap.2017.08.026

Morrison, L.K. (2004). Best practices in incident investigation in the chemical process industries with examples from the industry sector and specifically from Nova Chemicals, Journal of Hazardous Materials, Volume 111, 161–166.

Oswald, D, Sherratt, F., & Smith, S. (2018). Problems with safety observation reporting: A construction industry case study, Safety Science, Volume 107, 35–45. https://doi.org/10.1016/j.ssci.2018.04.004

Wincek, J.C. (2016). Amiss, a miss, a near miss. American Institute of Chemical Engineers 2015 Spring Meeting, 11th Global Congress on Process Safety, Austin, Texas, April 27–29, 2015, Process Safety Progress, Volume 35, No. 2, 137–142. https://doi.org/10.1002/prs.11757

Woźniak, Z. and Hoła, B. (2024). Analysing near-miss incidents in construction: A systematic literature review, Applied Sciences, Volume 14, 7260. https://doi.org/10.3390/app14167260