The Future of Crisis Response Training: AI-Generated Feedback for Incident Commanders?

Reet Kasepalu
Estonian Academy of Security Sciences
reet.kasepalu@sisekaitse.ee

Stella Polikarpus
Estonian Academy of Security Sciences
stella.polikarpus@sisekaitse.ee

Daniel Jefimov
Estonian Academy of Security Sciences
daniel.jefimov@sisekaitse.ee

Tambet Kütt
Estonian Academy of Security Sciences
tambet.kytt@sisekaitse.ee

ABSTRACT

In today's complex crisis landscape, effective incident command training relies on dynamic decision-making assessment frameworks. However, the Effective Command Behavioral Marker Framework (EC) momentarily demands a lot from human assessors, who must evaluate 72 criteria across a 5-point scale, leading to excessive cognitive load. This study explores whether AI-generated written feedback can support in the future assessors and possibly enhance learning outcomes by providing structured, data-driven insights of command training. We examine the assessment results from the solutions of 85 incident commanders to a virtual simulation "School Fire" scenario. Key challenges brought out in the feedback included delayed decision-making, inter-agency coordination gaps, and situational awareness deficits. While expert feedback is valuable, the time constraints for compiling the written feedback create a discrepancy in the length and quality of the provided feedback. This study explores how AI can complement human assessor's by reducing cognitive overload and enhancing incident command training through structured, data-driven feedback that supports assessors s expert judgment. Dynamic decision-making feedback systems using Human-AI collaboration may redefine training methodologies for next-generation incident commanders.

Keywords

Effective Command Behavioral Marker Framework (EC), Virtual simulation-based training (VSBT), AI, Human-AI collaboration.

INTRODUCTION

Structured and formative feedback is essential for developing the decision-making and situational awareness skills of incident commanders, who must operate under extreme pressure in high-stakes crisis scenarios. During Effective Command Behavioral Marker Framework based assessments assessors are responsible for three main activities. Firstly, they use virtual reality software to offer dynamic play-through of an assessment scenario. Then they construct an assessment interview with the incident commander. At last, they fill in the certificate in the Effective Command platform. Therefore, the assessment process itself imposes a significant cognitive load on the assessors (Bastian et al., 2024), who must evaluate each incident commanders performance across eight core behaviors, with nine sub-criteria per behavior, each rated on a five-point scale (Effective Command, 2025; Lamb et al., 2021) This complexity requires assessors to maintain a mental model of $8 \times 9 \times 5 = 360$ performance indicators, making it highly challenging to ensure consistency, specificity, and completeness in written feedback. To provide structured written feedback on the certificate immediately after the virtual simulation-based training

and interview makes the orchestration of the simulation-based training highly demanding (Prieto et al., 2018). Given these challenges, AI-assisted feedback generation presents a promising avenue for supporting assessors in delivering high-quality, structured, and actionable feedback that enhances the professional learning (Baker, 2016; Holstein et al., 2019).

Previous research on automated feedback generation (AFG) has primarily focused on educational settings, such as writing instruction, formative assessment in classrooms, or online learning environments. AFG could offer scalable, just-in-time adaptive feedback that addresses misconceptions and supports instructors—especially in contexts where personalized feedback is difficult to deliver manually (Song et al., 2023). However, few studies have explored the application of generative AI for high-stakes, simulation-based training, particularly in crisis management domains. Our study contributes to this emerging area by combining verbal, written, and performance data to support AI-generated feedback aligned with domain-specific behavioral criteria—marking a shift from content-focused educational feedback to performance-based situational assessment.

Research on feedback highlights its significant impact on learning and performance, with an average effect size of 0.79 (Hattie & Timperley, 2007), indicating a strong positive influence. In educational contexts, an effect size of 0.40 is considered a typical impact of schooling, suggesting that feedback is nearly twice as effective as standard instructional strategies. When applied to incident commanders, this finding underscores the critical role of structured feedback. The substantial effect size suggests that systematic and timely feedback mechanisms can accelerate the development of expertise, reduce errors in high-risk environments, and improve overall incident response efficiency. Given the dynamic and unpredictable nature of emergency management, integrating structured feedback into training and assessment procedures could provide a high-impact, evidence-based approach to improving command performance and crisis management outcomes. Effective feedback is a powerful tool for learning and performance improvement, characterized by several key attributes that enhance its impact. It must be goal-oriented, providing clarity on objectives and linking to real-world practice to ensure relevance (Bartlett et al., 2017; Hattie & Timperley, 2007). High-quality feedback is also specific and actionable, offering clear descriptions of strengths and areas for improvement while being linked to observable behaviors (Bartlett et al., 2017). To maximize its effectiveness, feedback should be timely and relevant, ensuring that learners can act on it when it is most useful. Additionally, limiting and prioritizing key points prevents cognitive overload, allowing learners to focus on the most critical areas for growth. Effective feedback supports self-regulation and deep learning by providing actionable strategies for improvement and encouraging deliberate practice, where learners repeatedly refine their skills through purposeful engagement. Finally, an optimal balance between positive reinforcement and constructive guidance ensures that feedback fosters both confidence and skill development, promoting continuous learning and expertise acquisition. These studies collectively highlight both the cognitive demands placed on assessors and the evidence-based benefits of structured, high-quality feedback, which our current work builds on by exploring how AI can help deliver such feedback more consistently and efficiently.

However, more research is required to support the work of assessors during assessments using Effective Command (Polikarpus et al., 2023). A recent study found that when the same incident commander's behaviors are assessed three times over a period using different virtual simulation scenarios, their three strongest behaviors were consistently Perception, Comprehension, and Plan (Polikarpus et al., 2024). Conversely, their weakest behaviors were Communication, Prediction, and Review. While quantitative performance metrics of incident commanders have been extensively analyzed in previous research (Polikarpus et al., 2020, 2024), the content and quality of written feedback remain an under-explored area. Providing structured written feedback in the medical profession has led to specialists improving their feedback writing skills (Bartlett et al., 2017). Despite its crucial role in guiding professional development, written feedback has received limited systematic investigation, leaving a gap in understanding how assessors construct meaningful evaluations. Given that written feedback generation is a labor-intensive and cognitively demanding process, requiring assessors to synthesize complex observations under time constraints, there is a growing need to explore AI-assisted approaches that could enhance efficiency, consistency, and feedback quality. This study addresses this gap by examining whether AI can serve as a support tool for assessors, facilitating structured and actionable written evaluations.

This paper investigates whether AI can assist assessors in providing effective written feedback that potentially facilitates meaningful skill improvement for incident commanders. Specifically, we explore whether AI-generated or AI-assisted feedback can match or enhance human-generated feedback in terms of specificity, actionability, and overall perceived quality. Recent research (Banihashem et al., 2024; Sutherland et al., 2023) states that AI generated feedback is seen as complementary rather than a replacement for human feedback as AI excels at

providing structural feedback but struggles with critical evaluation and deep contextual understanding. In addition to this, people engage differently with AI feedback, with some fully adopting it and others critically evaluating its usefulness. Our study is based on feedback data from 85 incident commanders responding to a standardized fire scenario in a virtual simulation environment. The aim of this study is to see can AI in collaboration with human provide written feedback to incident commanders dynamic decision-making skills to reduce the assessor's cognitive overload. To assess AI's role in feedback generation, we examine three key research questions: **RQ1** What characteristics define high-quality feedback in the dynamic decision-making assessment of incident commanders? **RQ2a** To what extent do expert raters perceive differences in feedback quality across human-generated, AI-generated, and hybrid feedback based on quantitative performance data? **RQ2b** How do expert raters evaluate the quality, learning potential, and content of AI-generated feedback produced from qualitative input such as assessor-trainee discussions?

By analyzing feedback across human, AI, and hybrid human-AI collaboration models for one virtual simulation scenario we were able to investigate dynamic decision-making in a specific context. Furthermore, we discuss the potential of feedback systems using AI support to reduce cognitive load, improve assessment consistency, and enhance reflective learning for incident commanders. Ultimately, this research contributes to the development of next-generation training methodologies that integrate AI to optimize crisis leadership assessment and skill development.

METHODOLOGY

This case-study looked at the assessment results of one virtual simulation scenario of rapid-fire development in school building. The dynamic decision-making behaviors of Estonian working rescue incident commanders were systematically assessed by two certified assessors per one commander. The virtual simulation "School Fire" was designed according to the Collaborative Authoring Process Model for Virtual Simulations framework (Polikarpus et al., 2021) by two certified assessors and validated by other two assessors, then introduced to all assessors before taken to use. The analysis sample was based exclusively on the formal assessment results of operational first-level rescue incident commanders recorded in the EC database (Polikarpus & Danilas, 2021). The training framework for these commanders and the evaluation of their behaviors through virtual simulations have been extensively documented, in prior studies (Polikarpus et al., 2020). The methodology for assessor training and certification is outlined in (Polikarpus & Danilas, 2021), whereas the comprehensive process of scenario development, validation, and the implementation of virtual simulation-based training is detailed in a doctoral dissertation (Polikarpus, 2024).

Data Collection

In Estonia, the Effective Command methodology involves two assessors evaluating each in-service incident commander and completing an assessment certificate (Polikarpus et al., 2020). This study analyzed first-level incident commanders' assessment results from the Effective Command platform from one scenario. The assessment focuses on eight behavioral aspects, including perception, comprehension, prediction, decision-making, planning, communication, command, and review, which are evaluated using nine criteria on a Likert scale (1–5). Each incident commander is evaluated with the overall assessment color (green, amber, or red) and written feedback. A green rating signifies that the incident commander demonstrates excellent dynamic decision-making competence. An amber rating indicates that the dynamic decision-making competence meets the occupational qualification standard, whereas a red rating reflects competencies below the required threshold (Polikarpus et al., 2020).

Scenario “School Fire”

The scenario that all 85 incident commanders had solved was a school basement fire scenario which is a high-stakes training case designed to assess first-level incident command competencies in a dynamic and constrained environment. The fire originates in a basement cloakroom due to electrical overloading, but the situation is complicated by a scheduled water supply outage caused by street construction, delaying the activation of the school's automatic fire suppression system. The fire spreads rapidly, fueled by high smoke accumulation and flashover appears in the basement. Eventually fire reaches the attic through construction scaffolding (see Fig. 1). With 62 students and five staff members present in school, evacuation becomes chaotic, as some students become trapped on the rooftop while others scatter in multiple directions, making search-and-rescue operations more complex. A security guard attempting to control the fire with an extinguisher is forced to take refuge in the sprinkler control room in basement. He is unable to escape from the room due to dense smoke and needs to be rescued. The incident commander arriving on scene must quickly assess the situation, manage resource

limitations, and determine appropriate suppression tactics while also coordinating with police, emergency medical services, and municipal authorities to restore water supply. The scenario evaluates key decision-making skills, including situational awareness, risk assessment, multi-agency coordination, and evacuation management. It provides a realistic and evidence-based framework for refining leadership performance in crisis situations, emphasizing the importance of dynamic risk assessment, procedural adherence, and strategic foresight in complex emergency management.

Figure 1. Bird-View of the Incident Site from the Assessor's Computer (Left) and the Incident Commander's View on Arrival in Virtual Reality (Right).

Process

An Excel file with two data columns (the anonymized ID and the written feedback) for 85 incident commanders was extracted from the dataset. To get an overview of which of the eight dynamic decision-making behaviors (perception, comprehension, prediction, decision-making, planning, communication, command, and review) had been addressed in the written feedback, a keyword-matching algorithm was implemented using OpenAI's GPT-4o model. This rule-based algorithm was developed in Python and used a custom-built dictionary of domain-specific keywords derived from the Effective Command assessment framework. Each feedback comment was analyzed for the presence of critical terms associated with the eight assessed behaviors (e.g., "risk assessment," "360-degree reconnaissance," "task delegation," "communication with dispatch," etc.). The algorithm performed case-insensitive matching and allowed for stemming and synonym recognition (e.g., "evacuate," "evacuation"). The frequency and co-occurrence of these keywords were used to cluster feedback around commonly identified problem areas, enabling a structured comparison across cases. In addition to the assessment criteria, we also presented the scenario description of the context to specify the keywords. The keywords were checked by an expert assessor. The keywords AI used for thematic analyzes in each 8 dynamic decision-making behavior can be found below.

Domain-Specific Keywords Used

1. Information Gathering

information gathering, initial information, reconnaissance, key personnel, building plan, documents, risks, hazards, visual information, questioning, drone, fire alarm system (FAS) schematics, 360-degree reconnaissance, information exchange, number of casualties

2. Understanding Information

incident size, scope, complexity, risk assessment, resources, specialized capabilities, hazard recognition, hazard zone, risk analysis, risk information, hazardous substances, fire compartments, casualty locations

3. Forecasting

incident progression, potential impact, broader impact, risk prevention, key indicators, priorities, risk evaluation, dynamic development, resource planning, fire development

4. Decision-Making

decision, strategy, risk analysis, time-critical decision, tactical choice, justified decision, life-saving, timely decision-making, inter-agency cooperation

5. Planning

incident resolution, tactical direction, action plan, resource allocation, tasks, priorities, safety measures, risk mitigation, collaboration, plan adjustment, occupational safety, tactics

6. Communication

communication, radio communication, briefing, commands, command channel, information transmission, clear messaging, cooperation partners, communication channels, feedback, risk communication

7. Incident Site Management

command structure, task allocation, determination of command span, effective action, inter-agency cooperation, management, operational structure, incident management

8. Review

incident evaluation, goal achievement, risk assessment update, resource requirement update, decision review, incident summary, tactical adjustment, plan revision

In addition to analyzing the feedback entries based on the eight behaviors, AI also checked whether the compulsory information had been added to the feedback. Namely, the feedback needs to be signed by both the assessor as well as the incident commander.

Generating a Structure for the Assessment

Momentarily, the structure of the written feedback to incident commanders is freeform and we wanted to see which structure could be suggested for future assessors. For this purpose, we created three different models: (1) a model proposed by an expert assessor, (2) a model that assessed all eight behaviors separately, and (3) a grouped model suggested by AI, developed after analyzing 85 feedback entries related to a single scenario and its assessment criteria. This resulted in 18 feedback entries being evaluated as the highest and second-highest scoring feedback entries in each model were extracted, and feedback that appeared in at least two models was blindly evaluated by an expert using a 0–5 Likert scale. The expert assessment focused on the extent to which the feedback aligned with the quantitative data across all eight dynamic decision-making behaviors, as well as its overall understandability.

Generating Criteria for High Quality Feedback

To generate criteria for high-quality feedback, feedback entries that had received high scores—either "brilliant" (5) or "very good" (4)—in the expert assessments were used as exemplars. These entries served as training input for the AI, from which we extracted key characteristics and patterns aligned with established criteria for effective feedback. Based on these inferred quality indicators, the AI was then prompted to generate improved versions of lower-scoring feedback entries. This process resulted in three distinct types of feedback for each case: (1) the original human-generated feedback, (2) AI-generated feedback based solely on the quantitative performance scores, and (3) AI-enhanced feedback where the original human-written text was modified considering both the assessment criteria and the quantitative data. To evaluate the perceived quality of each type, two independent experts performed a blind rating using a 0–5 Likert scale.

Generating Feedback Based on Transcribed Debriefing Discussions

An additional AI-feedback generation method was implemented to explore whether richer contextual input would enable the production of higher-quality written feedback. A staged and video-recorded formal assessment included reflections on decision-making, risk assessment, planning, and overall situational performance. The recording was transcribed using an Estonian speech-to-text tool (Olev & Alumäe, 2022), yielding verbatim textual representations of the post-simulation discussions. This transcription, along with the original scenario description and a set of predefined criteria for high-quality feedback, were then used as input for AI to generate written feedback for one case. This process resulted in a distinct type of feedback that leveraged naturalistic performance data and expert dialogue, allowing the AI to synthesize feedback that was grounded in both observable behaviors and expert interpretations. An expert assessor was also asked to write an assessment based on the audio discussion and video recorded simulation. To evaluate this new AI-generated feedback, a blind assessment was later conducted by an independent expert rater using a standardized 0–5 Likert scale. The feedback was rated across three dimensions: quality, learning potential, and content. This allowed for a systematic comparison between human-generated and AI-generated feedback based on transcribed debriefings.

Data Analysis

To explore the relationship between assessment result categories and the length of written feedback, a Spearman rank correlation analysis was performed. The ordinal variable representing result categories (red, amber, green)

was correlated with the word count of the corresponding feedback entries. This non-parametric statistical test was chosen due to the ordinal nature of the categorical variable and the potential for non-linear associations. All statistical tests and vizualisations were created using RStudio 2024.04.1.

To determine which model based on the eight behavioral domains would produce the most effective structure for high-quality feedback, we compared three different models named above. Each incident commander's performance was scored cumulatively across these domains according to the three models by a third expert assessor. The two highest-scoring feedback entries were extracted, and feedback that was present in at least two of the models was compiled for further evaluation, which resulted in analyzing 18 distinct feedback items. A first and second expert assessor then independently evaluated 18 different feedback entries using a 5-0 rating scale. The evaluation criteria included (a) coherence between the qualitative feedback and the quantitative scores provided by assessors, (b) the structural organization of the feedback, and (c) the clarity and comprehensibility of the written content. This approach allowed us to systematically compare the effectiveness of each model and assess which framework produced the highest-quality feedback for incident commander evaluations. Only high-level human-generated feedback was used to train AI.

RESULTS

RQ1: What Characteristics Define High-Quality Feedback in the Dynamic Decision-Making Assessment of Incident Commanders?

Based on the selected feedback from the dataset and additional expert-generated feedback, AI analyzed what makes a feedback entry "good" in the context of evaluating incident commanders. AI suggested that high-quality feedback is characterized by several key features.

- specificity and detail ensure that feedback provides clear, concrete examples of observed behaviors, describing what happened, what was done well, and what needs improvement.
- a balanced perspective acknowledges positive aspects before discussing issues, avoiding overly critical or vague statements.
- actionable insights differentiate high-quality feedback by offering solutions or next steps instead of merely pointing out problems.
- references to operational standards and best practices, comparing actions taken against established protocols or best practices.
- clarity and logical structure further enhance effectiveness by ensuring that feedback is well-organized, often structured by key evaluation criteria such as perception, decision-making, communication, and command. This makes it easy to follow what was done well, what went wrong, and why it matters.
- reflection and learning, strong feedback promotes self-assessment, helping recipients evaluate their decision-making processes.
- professional and constructive tone ensures that feedback remains formal yet supportive, maintaining a neutral and non-accusatory stance.

In addition to this, we also had a look at the length of the feedback. The results indicate that assessors provide more detailed feedback when the assessment outcome is under the threshold, while higher-performing incident commanders (green result) receive shorter feedback. A Spearman rank correlation analysis revealed a statistically significant negative correlation between assessment results and the word count of written feedback, $r_s = –0.36$, $p = 0.0008$. This suggests that as performance improves, the amount of written feedback decreases. This finding aligns with our objective of ensuring that those who require improvement receive more comprehensive feedback. Detailed feedback is crucial for incident commanders in the lower performance categories, as it helps them understand their shortcomings and provides guidance on how to address them effectively. Moving forward, maintaining this trend while ensuring that feedback remains structured and actionable will be key to supporting incident commander development.

When evaluating the 85 written feedback entries of incident commanders, this revealed both strengths and areas requiring significant improvement. While most feedback demonstrated adequate specificity and detail, providing clear descriptions of observed behaviors, other key quality criteria were inconsistently applied. Notably, references to operational standards and best practices were frequently absent, reducing the credibility and instructional value of the feedback. Although actionable insights were present in some entries, many feedback instances merely identified problems without providing concrete recommendations for improvement. Furthermore, while most feedback maintained a professional tone, the level of balanced perspective varied, with some entries focusing solely on deficiencies rather than offering a constructive mix of strengths and areas for growth.

RQ2a Differences in Feedback Quality across Human-Generated, AI-Generated, and Hybrid Feedback Based on Quantitative Performance Data

We first looked whether there are significant differences in feedback ratings across three types of feedback: **Human feedback, AI-generated feedback (AI),** and **Human-AI** collaborative feedback. This feedback was generated based on assessment data from the Effective Command framework eight behaviors, each with nine criteria on a scale 1-5. Two expert raters independently assessed the feedback using a 5-point scale. The results indicate that for Expert Rater 1, feedback from the Human–AI collaboration was rated significantly higher than both Human-only and AI-only feedback ($p < .05$), while no significant difference was found between AI and Human feedback (see Table 1). For Expert Rater 2, both Human and Human–AI feedback was rated significantly higher than AI feedback ($p < .05$), but there was no significant difference between Human and Human–AI.

Table 1. Descriptive Statistics of Expert Assessors Feedback Ratings on a Scale 0-5

Feedback Type	Expert Rater	Mean Score	Median Score	SD
AI	Expert Rater 2	1.17	1	0.41
AI	Expert Rater 1	1.50	1.5	0.55
Human	Expert Rater 2	2.17	2	0.41
Human	Expert Rater 1	2.00	2	0.00
Human & AI	Expert Rater 2	2.17	2	0.41
Human & AI	Expert Rater 1	3.33	3	1.03

These results suggest that feedback type significantly influences ratings, though the pattern of significance differs between expert raters. Expert Rater 1 showed a strong preference for Human-AI collaborative feedback, rating it significantly higher than both AI-only and human-only feedback. These findings suggest that hybrid AI-assisted feedback may be more effective for some evaluators than others, and future research should explore the conditions under which Human-AI collaboratively generated feedback is most beneficial. Boxplot in Figure 2 shows the distribution of scores for AI-generated, Human-generated, and Human-AI collaborative feedback. Human-AI feedback was rated significantly higher by Expert Rater 1, particularly in quality and coherence with assessment data, while AI-generated feedback was rated lowest.

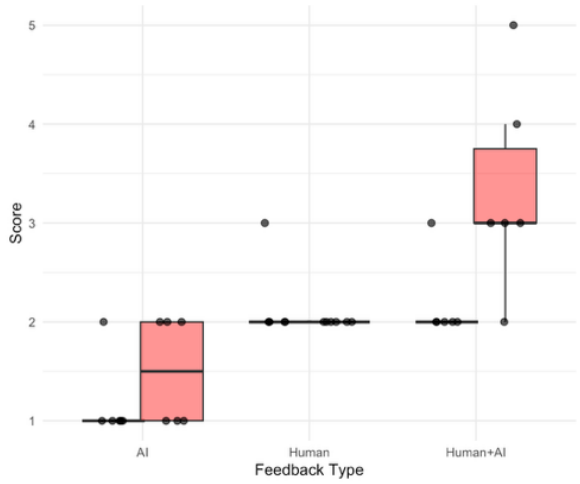

Figure 2. An Expert Rater Assessed Human-AI Collaboratively Generated Feedback Highest

The results indicate that AI-generated feedback (AI) received the lowest ratings, while Human-AI collaboration had the highest average score, particularly from Expert Rater 1 ($M = 3.33$, $SD = 1.03$). Expert Rater 2's assessments were more stable, showing less variation. A Kruskal–Wallis test was performed separately for each rater to determine whether the differences in ratings across the three feedback types were statistically significant. For Expert Rater 1's ratings, the Kruskal–Wallis test indicated a significant difference among the three feedback types, χ^2 (2) = 11.9, $p = 0.00254$. For Expert Rater 2's ratings, a significant difference was also found, χ^2 (2) = 11.2, $p = 0.00366$. Since the Kruskal–Wallis test was significant for both expert raters, post-hoc pairwise comparisons using Wilcoxon rank-sum tests with Bonferroni correction were conducted, which revealed further nuances. For Expert Rater 1, feedback from the Human-AI collaboration model was rated significantly higher than both AI-generated feedback ($p = 0.024$) and human-only feedback ($p = 0.027$) (see Table 2), while no significant difference was found between AI and human feedback ($p = 0.211$).

Table 2. Pairwise Comparisons of Three Tested Feedback Models Expert Ratings (Wilcoxon Test with Bonferroni Correction)

Expert Rater	Comparison	W Statistic	p-value
Expert Rater 1	AI vs Human	9.0	0.211 (ns)
Expert Rater 1	AI vs Human & AI	1.5	0.024 (*)
Expert Rater 1	Human vs Human & AI	3.0	0.027 (*)
Expert Rater 2	AI vs Human	2.5	0.023 (*)
Expert Rater 2	AI vs Human & AI	2.5	0.023 (*)
Expert Rater 2	Human vs Human & AI	18.0	1.000 (ns)

In contrast, Expert Rater 2 rated both human-only and Human-AI feedback significantly higher than AI-generated feedback ($p = 0.023$ for both comparisons). Nevertheless, no significant difference between human-only and Human-AI feedback ($p = 1.000$) was found. These results suggest that while the Human-AI model is generally perceived as superior to AI alone, the added value over human-only feedback may depend on the rater's preferences or evaluation style.

RQ2b: Evaluating the Quality, Learning Potential, and Content of AI-Generated Feedback Produced from Qualitative Input

The AI-generated feedback was produced using the speech-to-text transcription of the virtual simulation play-through, the subsequent discussion between two assessors and the incident commander, the scenario description, and predefined criteria for high-quality feedback. Unlike the earlier used approach, where AI alone was insufficient to generate high-quality feedback, the written text input data proved sufficient for producing feedback of notable quality. To assess the effectiveness of AI-generated feedback, we conducted a blind evaluation in which an expert rater compared AI-generated feedback to human-written feedback (written by an other expert assessor). The analysis focused on comprehensiveness, potential for supporting learning and quality. The results provide insights into the potential of AI-generated feedback in supporting expert assessments and its implications for training and evaluation practices.

After a blind expert assessment, quality received the highest average rating ($M = 4.22$, $SD = 0.67$) for AI generated feedback, indicating that the feedback was generally well-regarded in terms of correctness and effectiveness. The median was 4, with ratings ranging from 3 to 5. Learning potential had a slightly lower mean ($M = 3.78$, $SD = 0.83$) and a median of 4, suggesting that while the feedback was often seen as useful for learning, it showed more variation in quality. Content exhibited the widest spread ($M = 3.89$, $SD = 1.05$), with ratings spanning from 2 to 5, and a median of 4. The higher standard deviation suggests more variability in how well the feedback covered all necessary aspects. Overall, the results indicate that while the quality of feedback was consistently high, its learning potential and content were more variable, suggesting that some feedback provided clearer instructional value and broader coverage than others.

Table 3. Comparison of AI vs. Human Feedback Assessment

Variable	AI Mean	AI SD	AI Median	Human Mean	Human SD	Human Median
Quality	4.22	0.67	4	4.50	0.58	4.5
Learning Potential	3.78	0.83	4	4.50	0.58	4.5
Content	3.89	1.05	4	4.25	0.50	4.0

Human feedback received higher mean scores across all criteria. Quality ($M = 4.50$) was rated slightly higher than AI feedback ($M = 4.22$), indicating that expert-written assessments were perceived as more accurate and effective. Learning potential showed the largest gap, with human feedback ($M = 4.50$) notably higher than AI feedback ($M = 3.78$), suggesting that human-generated feedback was more instructional. Content was also rated higher for human feedback ($M = 4.25$) compared to AI ($M = 3.89$), implying that expert feedback covered more relevant aspects.

DISCUSSION

High-quality feedback in the evaluation of incident commanders is specific, structured, and actionable (Bartlett et al., 2017; Hattie & Timperley, 2007), providing both positive reinforcement and targeted suggestions for improvement. It follows a logical format, referencing best practices and operational standards, and encourages reflection and learning. Additionally, feedback length correlates negatively with performance, meaning that

weaker performers receive more detailed feedback, aligning with the goal of guiding improvement where it is most needed. Moving forward, assessors should prioritize integrating procedural references, structuring feedback to encourage self-reflection, and ensuring that all feedback includes clear, actionable recommendations to enhance the learning process for incident commanders. We suggest that having AI collaborating with the assessors while writing the written feedback making sure that the criteria have been followed (such as reminding them that they have mentioned problems without suggesting possible avenues of solutions) might increase the quality of the feedback for the incident commanders.

The results show that, how the feedback is generated (**Human; AI, Human-AI**) significantly influences expert ratings to it. AI-generated feedback was consistently rated the lowest and Human-AI collaborative feedback rated the highest when looking at the feedback generated based on the assessment data. While this study utilized GPT-4 for all feedback generation tasks due to its strong performance in natural language generation, future research could explore and compare alternative generative models—such as GPT-3.5, Claude, or domain-specific LLMs—to evaluate differences in feedback quality, domain alignment, and computational efficiency. The differences in ratings suggest that human involvement remains essential in providing high-quality feedback, though AI assistance may offer benefits when used collaboratively. This also coincides when we looked at the feedback generated based on speech to text data. Given these findings, a hybrid approach, where AI-generated feedback is refined by human assessors, may offer a viable solution to balancing efficiency and quality in feedback generation. However, there were inconsistencies between expert raters, indicating that the perceived value of AI-assisted feedback may vary across individuals as had also been suggested previously (Baker, 2016; Sutherland et al., 2023). In addition to this, there are several technical constraints that would hinder the AI and human complementarity within the assessment procedure of incident commanders. For instance, the data from Effective Command platform is not available for direct download before the certificate submission into the database. The written feedback, however, needs to be completed right after the play-through. When considering using the speech-to-text data as the basis for AI generated feedback, this procedure is automated momentarily for the Estonian language, nevertheless, the server is hosted by a university (Olev & Alumäe, 2022) and the processing time varies depending on the number of users. These findings highlight the need for further research to explore how AI can best support human assessors and whether structured human-AI workflows could enhance consistency in feedback evaluation.

Our results showed that AI-generated feedback was mostly perceived as correct and effective, yet it varied in its ability to offer meaningful instructional support. This confirms previous research that stresses the strength of AI being particularly effective in providing structural feedback but lacking the depth required for critical evaluation and nuanced contextual understanding (Banihashem et al., 2024; Sutherland et al., 2023). The study by Banihashem et al. (2024) highlighted that AI feedback tends to be descriptive, focusing on structural elements rather than deep critique. This aligns with our findings, where human-generated feedback was viewed as more insightful in identifying specific weaknesses and areas for improvement. The greatest distinction between AI and human feedback emerged in terms of instructional value, with human assessors providing more actionable and engaging critiques that encouraged deeper learning (see results to RQ2b). The comparison between our results and prior studies suggests that AI feedback is most effective when used as part of a hybrid model, where human expertise complements AI-generated insights. While AI can efficiently provide structured feedback, it lacks the depth and adaptability needed to fully support complex learning tasks. Moving forward, research should explore strategies for integrating AI and human feedback to enhance both consistency and instructional depth. Future systems could focus on refining AI's ability to offer context-aware, targeted suggestions, making it a more reliable tool in educational settings.

While this study provides valuable insights into the effectiveness of different feedback types, several limitations must be acknowledged. Ethical concerns regarding the transparency and fairness of AI-generated feedback need to be considered, as potential biases or inaccuracies could influence assessments, which nevertheless also occur when using human assessors. Technical constraints related to natural language processing accuracy, system reliability, and training data may have impacted the AI's ability to generate high-quality feedback comparable to human evaluators. Language issues further complicated the process, as variations in linguistic expression and AI's ability to interpret nuanced language may have led to discrepancies in feedback quality, particularly when translating from Estonian to English. These limitations highlight the need for further refinement in human-AI collaboration assessment methods and improved assessor support to ensure fair, effective, and scalable feedback mechanisms.

CONCLUSION

This study demonstrates that high-quality feedback in dynamic decision-making assessments of incident commanders must be specific, actionable, and structured to promote learning and reflection. Our findings highlight that while AI-generated feedback is perceived as generally accurate, it falls short in providing

instructional depth and contextual nuance compared to human feedback. However, when AI is used collaboratively with human assessors, the resulting feedback is rated significantly higher, suggesting that hybrid models can enhance both efficiency and quality. These benefits were especially evident when AI-generated feedback was based on rich, contextual input such as transcribed debriefings. Nonetheless, variation in expert ratings and technical limitations—such as access to assessment data and real-time processing constraints—underline the importance of further research and system development. Overall, human-AI collaboration holds promise as a scalable solution for improving feedback quality in simulation-based training, if implementation is context-sensitive and aligned with pedagogical standards.

ACKNOWLEDGEMENTS

We thank all the contributors to the virtual simulation scenarios authors and assessors of the rescue incident commanders, as well as all the commanders who have done their best to solve the incidents.

REFERENCES

Baker, R. S. (2016). Stupid Tutoring Systems, Intelligent Humans. *International Journal of Artificial Intelligence in Education*, *26*(2), 600–614. https://doi.org/10.1007/s40593-016-0105-0

Banihashem, S. K., Kerman, N. T., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: peer-generated or AI-generated feedback? *International Journal of Educational Technology in Higher Education*, *21*(1). https://doi.org/10.1186/s41239-024-00455-4

Bartlett, M., Crossley, J., & McKinley, R. (2017). Improving the quality of written feedback using written feedback. *Education for Primary Care*, *28*(1), 16–22. https://doi.org/10.1080/14739879.2016.1217171

Bastian, A., Kaiser, G., Meyer, D., & König, J. (2024). The Link Between Expertise, the Cognitive Demands of Teacher Noticing and, Experience in Teaching Mathematics in Secondary Schools. *International Journal of Science and Mathematics Education*, *22*(2), 257–282. https://doi.org/10.1007/s10763-023-10374-x

Effective Command. (2025). *Level One Effective Command Incident Form*. Effective Command. www.effectivecommand.org

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Holstein, K., McLaren, B. M., & Aleven, V. (2019). Co-Designing a Real-Time Classroom Orchestration Tool to Support Teacher–AI Complementarity. *Journal of Learning Analytics*, *6*(2), 27–52. https://doi.org/10.18608/jla.2019.62.3

Lamb, K., Farrow, M., Olymbios, C., Launder, D., & Greatbatch, I. (2021). Systematic incident command training and organisational competence. *International Journal of Emergency Services*, *10*(2), 222–234. https://doi.org/10.1108/IJES-05-2020-0029

Olev, A., & Alumäe, T. (2022). Estonian Speech Recognition and Transcription Editing Service. *Baltic Journal of Modern Computing*, *10*(3), 409–421. https://doi.org/10.22364/bjmc.2022.10.3.14

Polikarpus, S. (2024). *The Role of Trainers in Designing and Implementing Virtual Simulation-Based Training in Rescue Organisations* [PhD theses, Tallinn University]. https://www.etera.ee/zoom/202339/view

Polikarpus, S., & Danilas, K. (2021). EESTI PÄÄSTEMEESKONNA JUHTIDE VISIÕPPEPÕHISE HINDAMISE RAKENDAMINE JA TULEMUSED. *Turvalisuskompass*, 31–54.

Polikarpus, S., Kasepalu, R., & Sarmiento Marquez, M. E. (2024). From Dynamic Decision-making Assessments Using Virtual Simulation-based Training to Personally Targeted Training of Incident Commanders. *Information Technology in Disaster Risk Reduction*.

Polikarpus, S., Ley, T., & Poom-Valickis, K. (2020). Developing the Situational Awareness of Incident Commanders: Evaluating a Training Programme using a Virtual Simulation. *Proceedings Estonian Academy of Security Sciences*, *19*, 195–226. https://doi.org/https://doi.org/10.15158/fe4h-ch75

Polikarpus, S., Ley, T., & Poom-Valickis, K. (2021). Collaborative Authoring of Virtual Simulation Scenarios for Assessing Situational Awareness. *Proceedings of the 18th ISCRAM Conference*.

Polikarpus, S., Luik, P., Poom-Valickis, K., & Ley, T. (2023). The Role of Trainers in Implementing Virtual Simulation-based Training: Effects on Attitude and TPACK Knowledge. *Vocations and Learning*, *16*(3), 459–486. https://doi.org/10.1007/s12186-023-09322-1

Prieto, Sharma, K., Kidzinski, Ł., & Dillenbourg, P. (2018). Orchestration Load Indicators and Patterns: In-the-Wild Studies Using Mobile Eye-Tracking. *IEEE Transactions on Learning Technologies*, *11*(2), 216–229. https://doi.org/10.1109/TLT.2017.2690687

Song, Y., Jin, H.-Y., Pan, Z., & Cutumisu, M. (2023). A Systematic Review of Automated Feedback Generation in Empirical Educational Research. *International Society of the Learning Sciences*, *31*(1), 1895–1896. https://doi.org/10.1007/s40593-020-00222-2

Sutherland, S. C., Machado, T., Mahajan, S., Mohaddesi, O., Matuk, C., Smith, G., & Harteveld, C. (2023). Exploring the Role of AI-Generated Feedback Tangential to Learning Outcomes. *IEEE Conference on Computational Intelligence and Games, CIG*. https://doi.org/10.1109/CoG57401.2023.10333239