information

MDPI

Article

Improving Construction Site Safety with Large Language Models: A Performance Analysis

Concetta Manuela La Fata [1,*], Gianfranco Barone [2] and Marco Cammarata [2]

[1] Department of Engineering, University of Palermo, Viale delle Scienze, Building 8, 90128 Palermo, Italy
[2] TopNetwork S.p.A., Via Imperatore Federico, 90143 Palermo, Italy; gianfranco.barone@top-network.it (G.B.); marco.cammarata@top-network.it (M.C.)
* Correspondence: concettamanuela.lafata@unipa.it; Tel.: +39-09123861865

Abstract

Hazard recognition on construction sites is crucial for ensuring worker safety. Traditional methods widely rely on expert assessments, on-site inspections, and checklists, which can be time-consuming and susceptible to human error. The integration of multimodal Large Language Models (LLMs), such as GPT-based systems, offers a promising opportunity to overcome these limitations. Therefore, this study evaluates the effectiveness of GPT-4o in recognizing workplace hazards from image inputs, with a specific focus on construction sites. The results indicate that the model can serve as a valuable decision-support tool for safety professionals by providing scalable and real-time insights. However, the study also highlights key limitations, including the model's reliance on general visual features rather than domain-specific safety knowledge, and the continued need for human supervision. Additionally, ethical concerns, including bias in AI-generated hazard assessments, data privacy, and the risk of over-reliance on AI, must be carefully managed to ensure these tools contribute responsibly and effectively to proactive risk management strategies.

Keywords: construction safety; hazard; occupational health; generative AI; GPT-4o

1. Introduction

Nowadays, work-related accidents and illnesses continue to cause significant human and economic losses across all sectors. In Italy, this issue is particularly pressing. According to the Italian National Institute for Insurance against Accidents at Work (INAIL), 416,900 work-related accidents were reported in 2025. Excluding student-related cases, 792 of these were fatal up to December 2025 [1]. These statistics underscore the urgent need for innovative and proactive strategies to enhance workplace safety [2], especially in high-risk sectors such as construction [3]. At the European level, the construction sector recorded the highest number of workplace fatalities in 2023, accounting for nearly a quarter (24%) of all fatal workplace accidents across the EU (Figure 1) [4].

Within this context, Hazard Recognition and Risk Perception (HRRP) are fundamental for the development and implementation of effective safety strategies, as they enable the reduction in workers' exposure to risks [5–7]. Nevertheless, construction industry stakeholders (e.g., engineers, or managers) often face challenges in effective HRRP, primarily due to the lack of adequate skills [8,9]. Moreover, the sector is particularly susceptible to human errors, with over 80% of accidents attributed to workers' negligence or incompetence [3,10–12]. Despite individual factors such as experience, risk tolerance,

Academic Editors: Xin Lu, Jianhua Yang, Xiaoxia Li, Dehao Wu and Wei Wang

Received: 29 December 2025
Revised: 7 February 2026
Accepted: 14 February 2026
Published: 17 February 2026

emotional state, and task familiarity strongly shaping how workers recognize and respond to hazards, organizational compliance continues to be prioritized over hazards perception-aware tools to support workers.

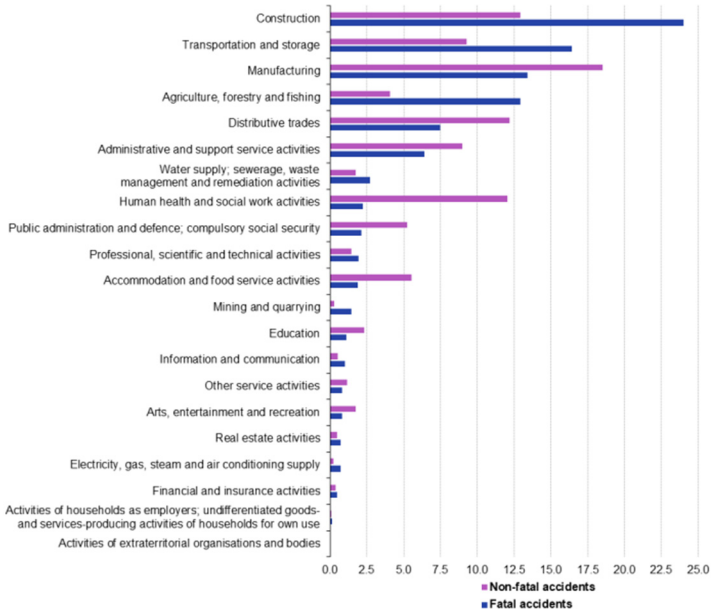

Figure 1. Percentage of fatal and non-fatal accidents at work, NACE section, EU, 2023 [4].

Given the pivotal role played by HRRP in construction safety management, a recent systematic review has traced their evolution over time [7]. Despite traditional methods such as Job Hazard Analysis (JHA) [13], on-site safety inspections, and checklists have been widely adopted, they strongly rely on human expertise, tend to be reactive rather than proactive, and often fail to detect non-obvious or emerging hazards, especially in fast-changing and dynamic environments [14,15]. In this context, emerging Industry 4.0 technologies offer significant potential for improving Occupational Health and Safety (OHS) in construction [16,17]. Among these, Virtual Reality (VR) and Augmented Reality (AR) platforms allow workers to experience high-risk scenarios in a controlled, risk-free environment, reinforcing visual attention and memory recall of hazards. Recent studies have also explored eye-tracking technologies to provide insights into how workers visually scan environments and why certain hazards are missed, thereby enabling the design of training programs tailored to individual attentional deficiencies [18,19]. Additionally, wearable physiological sensors are also being tested to measure stress, fatigue, and cognitive overload—factors closely associated with poor hazard recognition [20]. In parallel, Artificial Intelligence (AI) and its integration with other advanced technologies represents a transformative opportunity to automate and enhance real-time hazard monitoring [21]. Recently, AI has also been integrated with Building Information Modeling (BIM) to preemptively identify risks during the design and planning phases [22].

Despite their potential, the adoption of such technologies in construction faces several challenges, including high implementation costs, users' resistance, and the need for specialized expertise and on-site infrastructures, which may limit their accessibility especially to small and medium-sized enterprises [7]. In addition, they remain limited in replicating the full complexity and variability of construction sites and are primarily used off-site rather than for real-time, on-site hazard identification [23]. As a result, their usage is still largely confined to research settings, while their widespread adoption continues to be constrained by significant scalability challenges in everyday industry. Consequently, the construction sector is currently among the least digitalized globally [24].

These gaps create opportunities for complementary, low-cost, and accessible AI approaches—such as generative models like GPT—which may contribute to the development of more efficient and scalable automated hazard identification systems [25]. Developed by OpenAI, GPT is a Large Language Model (LLM) able to understand and generate human-like text [26]. Recent advancements have extended its capability to include multimodal functionalities, enabling the analysis of both textual and visual data [27–30]. Although still developing, GPT models have been extensively used in education for e.g., autodidactic experience and scientific writing [31], and healthcare for e.g., analyzing electronic records, preparing documentation, and diagnostic [32]. Similarly, in business and commerce, GPT models enhance customer interactions through chatbots and virtual assistants, optimize sentiment analysis, strengthen financial forecasting, detect fraud, and refine supply chain management [33]. On the other hand, studies on the usage of GPT models in the construction industry are still limited [30]. Nevertheless, these models hold significant potential in providing scalable, cost-effective solutions for continuous HRRP. In particular, their enhanced capability to analyze visual data may facilitate the detection of unsafe practices or conditions by processing images of workers, enabling real-time feedback and contextual recommendations that foster safer and proactive work environments while maintaining productivity.

Therefore, the primary objective of this study was to assess the effectiveness of GPT—specifically version GPT-4o—in hazard recognition within the construction sector. To guide this exploratory investigation, the following Research Question (RQ) was formulated: "How effectively can GPT-4o recognize hazards in construction-site images compared to independent OHS experts?" To this aim, the model's performance in hazard recognition was systematically evaluated using a set of uploaded images, and its assessments were compared with the corresponding evaluations provided independently by a team of experts in OHS. The findings not only highlight GPT-4o's potential as a decision support tool for workplace safety but also reveal its limitations and areas for improvement. Additionally, the results offer insights into strategies for enhancing the integration of AI technologies into safety monitoring systems, so improving hazard prevention and risk management.

The remainder of the manuscript is structured as follows. Section 2 reviews the existing contributions on the application of GPT models in the construction sector. Section 3 outlines the materials and methods adopted for data collection, image annotation, and comparative evaluation. Section 4 presents the results while Section 5 discusses their implications, highlighting both opportunities and limitations associated with the use of GPT-4o for construction safety. Finally, Conclusions are given in Section 6.

2. Related Work on GPT Models in the Construction Sector

To the best of the authors knowledge, only few contributions have explored the application of GPT models in the construction sector. In this regard, a literature review conducted by [30] examined the application of LLMs/generative AI—such as GPT models—in construction. The Authors observed that the adoption of these technologies is still low, with only five papers found at the time of their research. To address this gap, a team of experts was involved in complementing the existing studies and obtain deeper insights into the challenges and opportunities associated with the use of GPT models in construction. In ref. [3], the Authors demonstrated GPT's potential to support safety education for students preparing for a career in construction. Similarly, Ref. [34] investigated the use of ChatGPT 3.5 to support Construction Hazard Prevention through Design (CHPtD), an approach aimed to identify, eliminate, and manage potential hazards by integrating safety considerations into the design of construction projects. Students enrolled in an American civil engineering program were involved, and the study demonstrated that participants utilizing

ChatGPT during CHPtD sessions identified approximately 40% more hazards compared to those who did not use the AI tool. GPT's accuracy in construction projects was assessed in [35], interacting with the model via a questionnaire of thirty-six questions across four areas of risk management (i.e., identification, analysis, response, and monitoring). Results indicated moderate overall performance, with weaknesses in risk identification and prioritization due to limited real-world data access. In ref. [36], the Authors introduced a VR-based system integrating ChatGPT as a virtual instructor to enhance knowledge transfer and improve training outcomes. Involving fifty-two migrant workers having diverse linguistic and educational backgrounds, the designed system resulted in a 23% increase in knowledge scores post-training, proving its effectiveness especially for experienced workers than novices. Focusing on construction monitoring, Ref. [37] proposed an automated framework for generating daily construction reports from on-site videos by integrating ChatGPT 3.5 and computer vision methods. Implemented in a construction site in Hong Kong, the proposed framework reduced the time and effort needed for manual documentation while improving information sharing among project stakeholders. Ref. [38] explored the effective use of generative AI, particularly ChatGPT 4.0, in the Architecture, Engineering, and Construction (AEC) industry. The paper proposed practical guidelines to prompt engineering and design, aiming to optimize AI-driven applications for tasks such as construction scheduling, hazard recognition, and decision-making.

Recent studies have begun using multimodal GPT models on construction imagery. In ref. [39], a pre-trained Vision-Language Models (VLMs) (i.e., ChatGPT-4v's) was investigated for assessing its ability in describing images, with a focus on construction components, structural elements, and materials. In ref. [40], the Authors presented an end-to-end system that extracts visual features from a real-life construction site video and feeds them via structured prompts into a GPT-based reasoning module to identify hazards and support site safety.

3. Materials and Methods

A multi-phase approach was adopted to assess the performance of the selected generic LLM (i.e., GPT-4o) in hazards recognition within the construction sector (Figure 2).

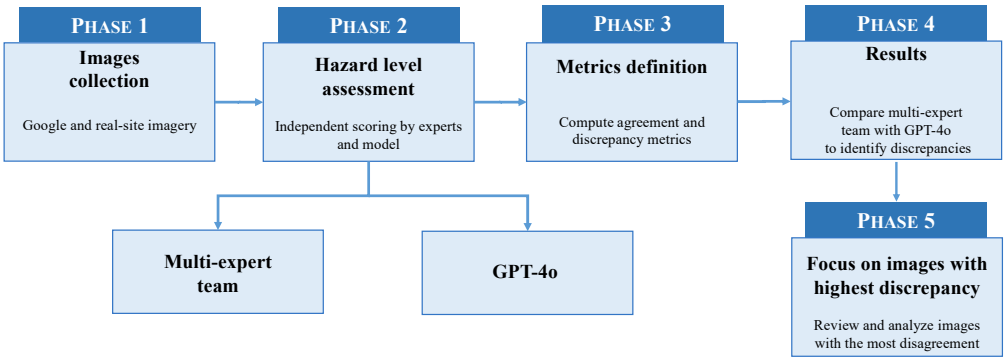

Figure 2. Multi-step methodology.

3.1. Images Collection

The dataset consisted of 102 images, systematically selected and categorized into two groups:

(a) Google images: this set consists of 51 creative commons license images sourced from Google, covering a wide range of work activities in construction sites (e.g., working at heights, operating heavy machinery, and handling hazardous materials). An example is provided in Figure 3.

(b) Images from real construction sites: this set consists of 51 images that are from different building construction sites in Sicily, capturing moments of daily work and workers' routine activities. An example is provided in Figure 4.

Figure 3. Example of Google image (Source: https://pixabay.com/photos/angle-grinder-work-sparks-skill-429757/, accessed on 13 May 2024).

Figure 4. Example of image from a real construction site.

3.2. Hazard Level Assessment: Experts vs. GPT-4o

HRRP in workplaces inevitably involves elements of subjectivity—influenced by hazard perception, training, and field experience—that can vary significantly among professionals. To address this issue, adopting a multi-expert approach is crucial for balancing subjective evaluations and ensuring a more robust and comprehensive response. This approach reduces individual bias and supports the identification of both immediate and emerging risks. With this recognition, a team of experts consisting of four professionals—each with 30 years of experience in the field of OHS—was involved in the present study to assess the aforementioned set of 102 images. Specifically, the team comprised three professionals from a Sicilian fire department and one labor inspector.

For every expert and image, the assessment was conducted synchronously, through an Excel worksheet. This approach allowed every expert to express his/her own judgment independently, without being influenced by the others. A four-level semantic scale (i.e., low, medium, high, and very high) was employed to rate the level of hazard, and every expert was asked to justify his/her own judgment with a brief explanation. During the assessment, several aspects were considered, with particular attention to:

(a) Immediate or potential hazards, such as checking for the presence of inadequate equipment, unsafe surfaces, or hazardous materials left uncontrolled.
(b) Correct use of Personal Protective Equipment (PPE), ensuring that tools such as helmets, gloves, protective goggles, and safety clothing were appropriately used.
(c) Compliance with safety regulations, verifying that all applicable laws and regulations related to the Italian construction sector and workplace safety were being followed.

The four-level semantic scale was then converted into a quantitative scale (Table 1) to allow for comparative analysis, as detailed in the following sub-sections.

Table 1. Semantic and quantitative scale.

Semantic Hazard Level	Quantitative Value
Low (L)	0.25
Medium (M)	0.5
High (H)	0.75
Very High (VH)	1

The dataset of 102 images was also analyzed using GPT-4o (release date 13 May 2024). Specifically, for every image uploaded, a new conversation was initiated, and the same prompt was used for every request. This ensured uniformity, independence and consistency in the results produced. Specifically, two distinct prompts were inserted. The first prompt was: "You are an expert evaluator of workplace safety. Based on the provided image, define the safety level with a value between 0 and 1, indicating 0 for minimal hazard and 1 for maximum hazard. In your evaluation, consider the situational context, its dynamics, applicable safety regulations, and the presence of adequate PPE". Differently from experts—who are more likely able to express their judgements by a semantic scale—allowing the model to use a continuous scale (ranging from 0 to 1) enabled a more nuanced assessment of hazard levels, capturing subtle variations that might be overlooked in more rigid analyses based on discrete-scale evaluations. Afterwards, a second prompt was used to request an explanation for the numerical value provided: "Briefly describe the reasons that led you to assign the given value". This allowed for a direct comparison between GPT-4o and expert evaluations.

In this study, a single-inference protocol was intentionally adopted to simulate the expected real-world deployment scenario of a GPT-based safety decision-support tool on construction sites. In operational contexts, hazard assessments must be generated in

real time from a single visual observation, under time, computational, and connectivity constraints, rather than through repeated sampling and consensus aggregation. Therefore, the methodological objective was to evaluate the model's behavior under realistic use conditions rather than under laboratory-style robustness testing designed to minimize stochastic effects. While this design does not explicitly quantify per-image output variance, the analysis focuses on aggregate performance patterns across a dataset of 102 construction-site images. At this scale, unsystematic stochastic fluctuations are expected to partially average out, reducing their influence on dataset-level findings. This approach is consistent with established practice in safety and risk assessment research, where observer variability is inherent, but conclusions are drawn from performance trends across many independent cases rather than repeated judgments of the same scene.

Expert and model judgements are detailed in Appendix A.

3.3. Metrics Definition

The following definitions are introduced [41].

(a) False Positive (FP): also known as type I error, it occurs when a test incorrectly rejects a null hypothesis that is true. It happens when the test finds evidence of something that does not exist.
(b) False Negative (FN): also known as type II error, it occurs when a test fails to detect a condition or phenomenon that is present. Type II errors are more concerning in safety contexts, as they represent situations where hazards are not detected by the model.
(c) True Positive (TP): the model correctly predicts the positive class, namely it identifies the presence of hazards.
(d) True Negative (TN): the model correctly predicts the negative class, namely it identifies the absence of hazards.

To evaluate the model's performance in identifying hazardous situations, a confusion matrix was developed using a binary classification based on a threshold value T^*. The threshold T^* represents the minimum score at which a situation is categorized as hazardous. It was established in consultation with domain experts, who defined T^* as the minimum intervention value. In practice, any model score higher than or equal to T^* indicates a hazardous situation requiring intervention. Conversely, scores below T^* indicate situations judged by experts as non-hazardous and not requiring intervention. The definition of the intervention threshold level is necessitated by the inherent fuzziness of human risk perception and the need to translate expert judgment into a binary decision for the model evaluation. Human assessment often operates with overlapping semantic categories of hazard or risk. For instance, a situation is not immediately classified as "Medium" at a score of 0.5; instead, its degree of membership in the "Medium/Actionable" category increases gradually as the mean expert score rises from 0.25 ("Low"). In this context, the threshold is not simply an average but rather a predefined, necessary boundary for converting a fuzzy, nuanced hazard assessment into the crisp "go/no-go" decision required for validating an automated hazard detection system. It formalizes the domain experts' consensus on the minimum level of hazard that justifies intervention. The ground truth was established by averaging the scores assigned by the team of experts involved. Therefore, a situation was classified as a true positive (TP) if both the model's prediction and the average expert score were equal to or higher than T^*, indicating the correct identification of a hazardous situation. A true negative (TN) was recorded when both the model and the average expert score were below T^*, indicating the correct identification of a non-hazardous situation. False positives (FP) were identified when the model's prediction was T^* or higher (indicating a hazardous situation), while the average expert score was below T^* (indicating a non-hazardous situation). Conversely, false negatives (FN) were identified when the model's

prediction was below T^* (indicating a non- hazardous situation), while the average expert score was T^* or higher (indicating a hazardous situation).

Therefore, the comparative analysis between experts and GPT-4o was conducted through the following metrics.

(a) Mean Absolute Error (MAE): it quantifies the average deviation of the model's predictions from the actual observed values. For n samples, if y_i is the actual value and $\hat{y}_i$ is the predicted one, MAE is calculated as follows (Equation (1)).

$$\text{MAE} = \frac{1}{n} \cdot \sum_{i=1}^{n} |y_i - \hat{y}_i|, \tag{1}$$

(b) Accuracy (A): it represents the number of correct predictions (i.e., TP and TN) out of the total number of predictions made, namely it indicates the percentage of cases where the model correctly classifies a situation in line with the expert judgments (Equation (2)).

$$\text{A} = (\text{TP} + \text{TN})/(\text{FP} + \text{TP} + \text{FN} + \text{TN}), \tag{2}$$

(c) Sensitivity (S): it measures the ability of a statistical test or classification model to correctly identify TP. In other words, it measures the proportion of TP identified out of the total number of TP and FN (Equation (3)).

$$\text{S} = \text{TP}/(\text{FN} + \text{TP}), \tag{3}$$

(d) Precision (P): it is a performance metric used to measure the accuracy of positive predictions in a classification problem. It quantifies the proportion of correctly predicted positive instances out of all instances classified as positive by the model (Equation (4)).

$$\text{P} = \text{TP}/(\text{FP} + \text{TP}), \tag{4}$$

(e) F1 Score (F1): it is a performance metric that combines both Precision (P) and Sensitivity (S) into a single measure, providing a balanced evaluation of the model's ability to correctly identify positive instances while limiting false positives. F1 score is defined as the harmonic mean of P and S, which makes it particularly suitable in cases where the class distribution is imbalanced or when both false positives and false negatives carry significant consequences. It is computed as follows (Equation (5)):

$$\text{F1} = (2 \times \text{P} \times \text{S})/(\text{P} + \text{S}), \tag{5}$$

4. Results

The metrics described in Section 3.3 were initially computed for every image by comparing the GPT-4o score with the arithmetic mean of the experts' individual evaluations (see Appendix A). Based on risk management considerations, T^* was set equal to 0.4 as explained in Section 3.3, justifying the need for appropriate corrective interventions. Afterwards, the resulting metrics were further aggregated for the Google and real images datasets separately, as well as overall (Tables 2–4).

Table 2. False positives and negatives for the two datasets.

Images	FP	FN
Google	2	16
Real	2	12

Table 3. Confusion matrix on both datasets.

	AI Predicts Hazardous Conditions	**AI Predicts Non-Hazardous Conditions**
Experts predict hazardous conditions	64 (TP)	28 (FN)
Experts predict non-hazardous conditions	4 (FP)	6 (TN)

Table 4. MAE, A, S, P and F1 values for the two datasets.

Images	**MAE**	**A**	**S**	**P**	**F1**
Google	0.23603	0.64706	0.66667	0.941176	0.871428
Real	0.18333	0.72549	0.72727	0.941176	0.820513
Both	0.20968	0.68627	0.696	0.941	0.8

The results obtained indicate as follows.

- MAE results indicate that GPT-4o produced more accurate predictions on real images than on Google images. The aggregated MAE value is 0.20968, indicating an overall average error of approximately 21% between the system's predictions and the evaluators' assessments.
- Sensitivity: the model performs overall well in identifying hazardous situations (~69.6%) but missed about 30.4% (false negatives).
- Accuracy (~68.6%): roughly 68.6% of the model's predictions matched expert consensus. However, about 31.4% were misclassified due to either FP or FN. Specifically, Table 2 shows that the model produced two FP for both the real and Google image datasets, indicating the presence of hazards that were not identified by the experts. As concerns FN, the model recorded 12 and 16 cases for real and Google images respectively. This suggests that the model seems generally less effective at detecting hazards in images sourced from Google.
- Precision: the model tends to be highly accurate in predicting hazardous conditions, with a precision of 94.1%.
- According to expert opinions, approximately 90% of the 102 images depict hazardous situations (i.e., TP + FN), suggesting a potential bias toward medium-high risk cases. This may reflect a tendency among individuals to overestimate risks, or conversely, that the abundance of safety protocols and regulations makes full compliance challenging. However, this strong imbalance in the dataset—where hazardous cases largely outnumber safe ones—requires careful consideration when evaluating the model's performance. Traditional accuracy metrics can be misleading in such contexts, as a classifier might achieve high accuracy simply by predicting the majority class more frequently. In this regard, the F1 score becomes a more reliable indicator, as it balances precision and recall, providing insights into the model's ability to correctly identify both positive and negative classes under class imbalance. The F1 score obtained for the Google images dataset is 0.871428, while it is 0.820513 for the real-image dataset. Both values indicate relatively strong performance, but the drop in the real-image score suggests that the classifier struggles more when compared with images closer to real-world conditions, where variability and noise are higher. Overall, these results imply that while the model generalizes reasonably well, the imbalance of the dataset may inflate the perceived performance.

4.1. Sensitivity Analysis of the Decision Threshold Under Class Imbalance

To investigate the impact of the decision threshold T^* on the model's performance, a sensitivity analysis was conducted by varying T^* from 0.1 to 0.9 and re-computing sensitivity (recall), precision, and accuracy values. The results are summarized in Figure 5.

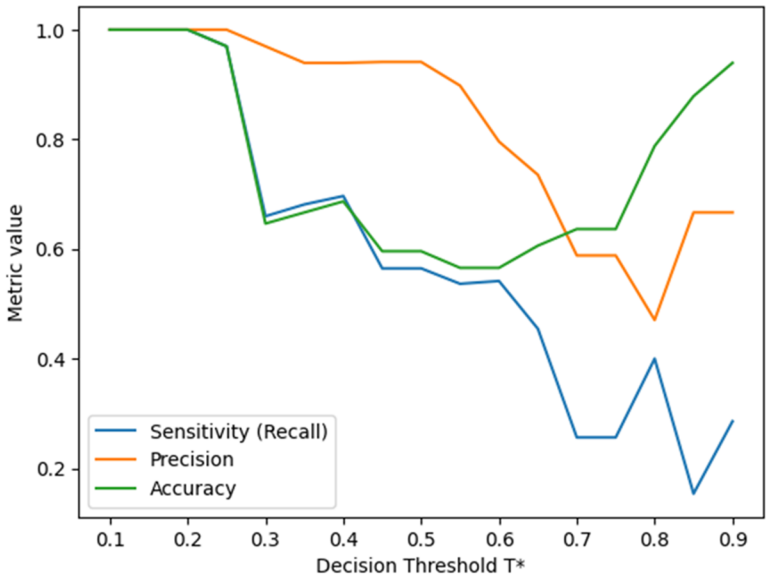

Figure 5. Sensitivity analysis results.

As shown in Figure 5, the choice of T^* has a clear and non-negligible influence on all evaluated metrics. This effect must be interpreted in light of the strong class imbalance of the dataset, in which hazardous situations (positive class) largely outnumber non-hazardous ones according to expert consensus.

At low threshold values ($T^* \leq 0.2$), sensitivity is close to 1.0, indicating that almost all expert-labeled hazardous situations are correctly identified by the model. This behavior is consistent with the dataset imbalance: when the threshold is very permissive, the model tends to classify most cases as hazardous, which favors the majority class and leads to very few false negatives. However, in this range, accuracy does not provide meaningful discrimination power, as correct classification of the dominant class drives the metric.

As the threshold increases beyond 0.3, sensitivity exhibits a marked decline, reaching values below 0.6 for $T^* \geq 0.5$ and dropping sharply for higher thresholds. This trend reflects a rapid increase in false negatives, which is particularly critical in safety-related applications. In an imbalanced dataset, even moderate increases in the decision threshold can disproportionately affect sensitivity, as hazardous cases dominate the ground truth.

Precision shows a different pattern. It remains high at low and intermediate threshold values and decreases more gradually as T^* increases. This behavior is again influenced by class imbalance: because hazardous cases are prevalent, predictions labeled as hazardous are often correct, which sustains high precision across a broad threshold range. The observed dip in precision at higher thresholds reflects the increasing scarcity of positive predictions and a less favorable balance between true positives and false positives.

Accuracy follows a non-monotonic trend and varies substantially across thresholds. Its relatively high values at both low and high thresholds illustrate a well-known limitation of accuracy under class imbalance: the metric can appear favorable even when the model's ability to detect hazardous situations (sensitivity) is significantly degraded. Therefore, accuracy alone is insufficient for assessing model performance in this context.

Importantly, the selected threshold T^* (i.e., 0.4) lies in a region where sensitivity, precision, and accuracy reach a reasonable compromise. At this value, sensitivity remains around

0.65–0.70, while precision stays high, limiting false alarms without excessively increasing missed hazardous cases. This balance is particularly relevant for occupational safety applications, where false negatives carry more severe consequences than false positives.

Overall, the sensitivity analysis confirms that the decision threshold substantially affects the confusion matrix and derived metrics, and that these effects are amplified by the unbalanced nature of the dataset. The results justify the explicit reporting and discussion of threshold selection and support the use of sensitivity and F1-based evaluations when validating AI-assisted hazard recognition systems in safety-critical and imbalanced settings.

4.2. Analisys of Images with High Discrepancy

Examining images where expert evaluations significantly differ from those provided by GPT-4o—measured by MAE values—is essential for identifying discrepancies and potential areas for improvement. These differences may reveal model interpretation errors e.g., stemming from training on a non-representative set of images, also disregarding external historical, cultural, and environmental factors. On the other hand, experts might be influenced by personal experiences, fatigue, or stress, which may lead to judgment errors or over/underestimations of risks.

Among images that exhibited the highest values of MAE, the one reported in Figure 6—file 15 in Table A1 provided in the Appendix A—is from Google dataset. According to expert evaluations, the mean hazard level was 0.8125, while the model assigned a value equal to 0.3. A brief overview of comments provided by both the experts and the model are given in Table 5.

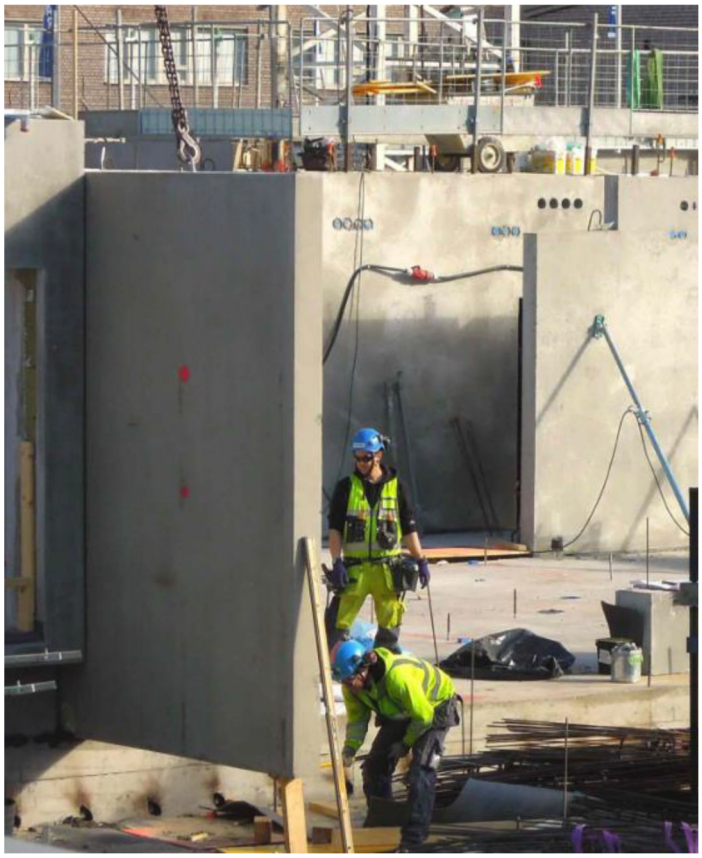

Figure 6. Example of Google image with high MAE (Source: https://pxhere.com/it/photo/644934, accessed on 13 May 2024).

Table 5. Comparison of human expert opinions versus AI model for the Google image.

Team of Experts	GPT-4o
Expert 1: Workers are properly equipped with all PPE but are operating within the suspended load area, exposing them to the risk of overturning. Expert 2: Operators are within the crane's operating area, near a suspended load. Expert 3: Presence of workers within the load handling area of a crane, wearing appropriate PPE. Expert 4: Workers are engaged in the placement of the precast concrete wall, within the crane's operating area. They are potentially exposed to crushing hazards due to hook and chain failure and possible slipping of the wooden block placed at the base of the precast structure.	Presence of PPE: workers are wearing safety helmets, high-visibility vests, and gloves, which significantly reduces the risk of injuries. Site conditions: the area appears well-organized and clean, with no obvious obstacles or clutter that could cause accidents. Safety systems: anchors and supports for lifting the panels are visible, indicating that safe lifting practices are being followed. Safety distance: workers maintain an appropriate distance from the most hazardous areas, such as the lifting point of the precast panels.

Analyzing the two evaluations, it is evident that experts possess a comprehensive understanding of regulations, technical requirements, and potential hazards that are not always evident through superficial observation. For example, the failure of the crane hook or the potential for a wooden block to slip—due to e.g., subtle ground slopes, micro-vibrations, or suboptimal support positioning—might seem unlikely events in a well-organized context and their identification requires a nuanced evaluation that only experts can provide. In contrast, GPT-4o primarily relies on the visual information available. As a result, the model seems unable to infer the risks that professionals can suggest from their extensive experience and the detailed analysis of each phase of the work process. Consequently, the model tends to emphasize reassuring elements—such as the presence of PPE—while overlooking less obvious yet potentially hazardous aspects (i.e., false negative cases). Moreover, GPT-4o does not account for industry-specific regulations, leading to an analysis that is more superficial, optimistic, and limited to the explicitly provided information, as aforementioned.

The image in Figure 7 is from the real images' dataset—file 42 in Table A2 provided in the Appendix A. The experts indicated an average hazard level of 0.6875. In contrast, GPT-4o provided a value equal to 0.3. A brief overview of comments provided by both the experts and the model are given in Table 6.

The experts' evaluation primarily focuses on technical aspects and critical issues related to the absence of PPE. They clearly emphasize that the workers are not wearing helmets or gloves and are positioned within the operating area of the concrete pump. Additionally, they highlight the concrete risk of crushing, not only from the proximity to the mechanical arm but also from the potential structural failure of the arm itself. The model, while acknowledging the importance of PPE, merely notes that the workers are at least wearing high-visibility vests, and that the construction site appears to have a certain level of organization (such as fencing and designated work areas). Unlike the experts, it does not explicitly confirm whether helmets, gloves, or safety glasses are being used, as these elements are not clearly visible. Instead, GPT-4o provides a more general assessment, recognizing both positive aspects and risks, particularly the likelihood of crushing and accidental contact with the concrete pump. These differences in emphasis are also reflected in the corresponding evaluations.

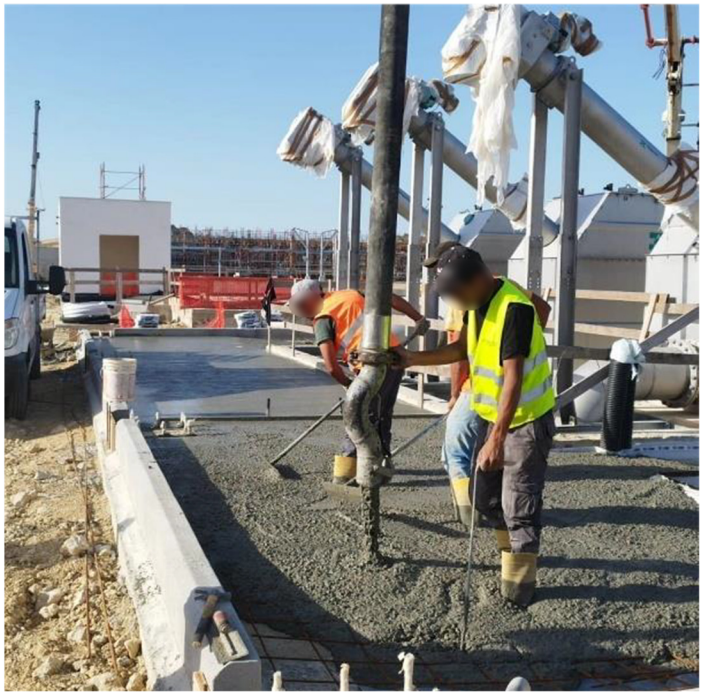

Figure 7. Example of real image with high MAE.

Table 6. Comparison of human expert opinions versus AI model for the real image.

Team of Experts	GPT-4o
Expert 1: Workers are not wearing helmets and are positioned within the operating area of the concrete pump. Expert 2: Operators are near the concrete pump discharge, without appropriate PPE. Expert 3: Work is being carried out within the operational area of mechanical arms, with workers lacking proper PPE. Expert 4: The two workers are not wearing essential PPE, such as helmets and gloves. They are positioned within the operating area of a concrete pump, posing a potential risk of crushing due to the possible failure of the pump arm.	Workers are wearing some PPE, such as high-visibility vests, which enhance their visibility. The construction site appears to be somewhat organized, with fencing and designated work areas. However, there are several concerns: it is unclear whether the workers are wearing protective helmets, safety glasses, or gloves; the work area seems to be exposed to other ongoing construction activities that may introduce additional hazards; the workers' proximity to the concrete pump increases the risk of accidental contact or crushing.

5. Discussion

5.1. Use of GPT-4o in Workplace Safety: Opportunities and Challenges

The integration of AI-based systems such as multimodal LLMs into workplace safety represents a promising opportunity [30], offering significant advantages while also raising critical concerns that warrant careful examination.

One of GPT-4o's key strengths lies in its ability to rapidly process and analyze large volumes of data, including images and textual descriptions. This capability makes it particularly valuable in complex work environments such as construction sites, where the timely identification of potential hazards is crucial. Further benefits are listed below:

- Accessibility and availability: multimodal GPT-4o can be available 24/7, providing instant access to hazard recognition and risk assessment guidance and information, which can be crucial in time-sensitive situations.

- Consistency: GPT-4o can provide consistent evaluations based on predefined criteria by design using a correct set of parameters, reducing variability that might occur with human judgment. The system also provides a detailed description of the identified hazard level. This allows for the analysis of the system's errors. By identifying critical points, its use could be limited to situations where it demonstrates high reliability.
- Scalability: it can handle multiple queries via API, making it scalable for large construction projects.

Despite these strengths, GPT-4o's analysis also reveals significant limitations. One of the primary concerns is its reliance on visible data and explicitly provided information, if available. Unlike experienced human experts, the model is unable to infer less obvious hazards that may not be immediately apparent. This limitation may result in false negative cases, where actual hazards remain undetected. This is particularly concerning, as failing to detect real risks in workplace settings can have severe consequences for worker safety. Another notable limitation is GPT-4o's lack of specialized regulatory safety knowledge while possessing an extensive generic knowledge across various domains [36]. Therefore, the model can process and analyze data but does not inherently apply specific workplace safety regulations. This shortcoming is evident in cases where the model underestimates critical hazards. Moreover, GPT-4o tends to prioritize visually reassuring elements—such as the presence of PPE or an orderly work environment—while potentially overlooking underlying hazards. While this approach may be useful for preliminary screenings, it is insufficient for conducting thorough and complex risk evaluations.

5.2. Ethical Considerations

The integration of GPT models into workplace safety raises fundamental ethical concerns regarding accountability, reliability, and the broader implications of delegating safety-related issues to AI. Over-reliance on AI could diminish human vigilance, increasing the likelihood of critical safety oversights. Instead, the present study underscores the necessity of using AI as a decision-support tool rather than a standalone safety assessment system, augmenting rather than replacing human expertise. Safety professionals still bring essential experience, regulatory knowledge, and contextual awareness that AI—regardless of its sophistication—cannot fully replicate.

Another significant ethical concern pertains to bias in AI-generated hazard recognition and risk assessment. GPT-4o is trained on pre-existing datasets, which may include non-representative or inaccurately labeled data. Consequently, the model's evaluations may be skewed, particularly in contexts with distinct cultural or technical considerations. For instance, if GPT-4o has not been trained on images depicting specific workplace hazards, it may fail to recognize them. This raises a crucial question: how can fairness and accuracy in AI-based assessments be ensured? Is it ethical to depend on a system that may inadvertently reinforce or amplify biases present in its training data? A potential solution lies in continuously updating the model with more representative datasets and involving human safety experts more extensively in the AI training process to improve its contextual awareness.

Privacy and data security also represent a critical ethical challenge. Images and textual descriptions processed by GPT models may contain sensitive information regarding workers, workplaces, or corporate environments. In real-world applications, sharing workplace imagery with an external AI system could raise concerns about confidentiality, proprietary information, and corporate security. Therefore, in alignment with the EU AI Act [42], it is essential to implement strict data protection policies, ensuring compliance with privacy laws and guarantee anonymity.

Finally, the social and economic implications of AI adoption in workplace safety must be carefully considered. On the one hand, AI-driven systems like GPT models can enhance efficiency and potentially reduce costs. On the other hand, they contribute to the fear of job displacement among human safety professionals. Therefore, organizations must strike a balance between leveraging the benefits of AI automation and fulfilling their responsibility to preserve employment opportunities. Investing in worker training and professional development remains essential to ensure that technological advancements do not come at the expense of human expertise.

6. Conclusions

This study represents a preliminary, exploratory performance analysis of a generic multimodal LLM (i.e., GPT-4o) applied to hazard recognition in construction, using image inputs and without domain-specific fine-tuning. The primary scientific contribution does not lie in achieving state-of-the-art predictive performance, but in systematically characterizing the strengths, limitations, and error patterns of a general-purpose LLM when compared against expert judgment in a safety-critical domain.

With an overall accuracy of approximately 69% and sensitivity near 70%, the model shows a moderate capability in identifying hazardous situations, even without domain-specific fine-tuning. As a result, the findings suggest that AI-driven systems can serve as a valuable tool to support human safety inspectors and professionals, offering scalable and real-time insights for improved workplace safety.

However, several critical challenges remain. The model's reliance on general visual features rather than domain-specific knowledge can lead to misclassifications, especially in complex scenarios. Additionally, factors such as lighting variations, occlusions, image quality, and country-specific safety regulations may affect performance, highlighting the need for more robust preprocessing techniques. Another key consideration is the interpretability of the textual justifications provided by the model, which should be evaluated against industry safety standards to ensure alignment with expert assessments. As a result, future research may focus on fine-tuning a multimodal LLM using curated datasets of construction site images annotated by safety experts, thereby improving both accuracy and sensitivity. Additionally, integrating multimodal data sources—such as IoT sensors and contextual site information—could further provide a more comprehensive risk evaluation. Moreover, refining the model's explainability will be essential to ensure that its assessments and justifications align with regulatory and practical safety requirements.

Limitations of the study must also be acknowledged. Despite the sample size of real images is sufficient for an exploratory analysis and for revealing indicative trends, it remains limited in both size and diversity and exhibits an inherent imbalance toward hazardous situations. Therefore, future research should employ a more extensive and heterogeneous dataset to establish a stronger empirical basis for evaluating the model. In addition, multimodal LLMs may produce slightly different outputs across repeated runs with identical inputs. Conversely, a single-inference protocol was intentionally adopted in the present study to evaluate the model's behavior. While this approach reflects the expected field deployment, future developments should complement it with controlled repeated-inference experiments to formally estimate prediction uncertainty and assess stability at the individual-image level.

By addressing these areas, AI-based risk assessment tools could become a key component of proactive HRRP strategies in the construction industry, contributing to reduced accidents and enhanced site safety. Crucially, AI should augment rather than replace human expertise, and safety professionals remain indispensable due to their contextual awareness, regulatory knowledge, and experiential judgment. Therefore, a balanced ap-

proach is essential, managing the interplay between AI efficiency and human judgment, ensuring safety, fairness, data privacy, and mitigating bias.

Author Contributions: Conceptualization, C.M.L.F.; Methodology, C.M.L.F. and G.B.; Data curation, C.M.L.F.; Writing—original draft, C.M.L.F.; Writing—review and editing, C.M.L.F., G.B. and M.C.; Software, G.B.; Formal analysis, G.B.; Supervision, M.C.; Validation, M.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: The methodological approach was designed in collaboration with the Research and Development (R&D) division of TopNetwork S.p.A., a company specialized in IT and digital transformation with a particular focus on artificial intelligence. The authors acknowledge Vito Pisciotta and Giovanni Lo Bianco for their assistance in data collection.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

LLM	Large Language Model
AI	Artificial Intelligence
HRRP	Hazard Recognition and Risk Perception
JHA	Job Hazard Analysis
OHS	Occupational Health and Safety
VR	Virtual Reality
AR	Augmented Reality
BIM	Building Information Modeling
CHPtD	Construction Hazard Prevention through Design
AEC	Architecture, Engineering, and Construction
PPE	Personal Protective Equipment
FP	False Positive
FN	False Negative
TP	True Positive
TN	True Negative
MAE	Mean Absolute Error

Appendix A

Table A1. Rating of Google images.

File ID	Expert 1	Expert 2	Expert 3	Expert 4	Mean Value of Experts' Judgement	AI Model
1	H	VH	H	H	0.8125	0.3
2	H	H	M	L	0.5625	0.3
3	M	H	M	H	0.625	0.2
4	H	VH	H	M	0.75	0.3
5	VH	VH	H	H	0.875	0.7
6	H	VH	H	H	0.8125	0.7
7	VH	VH	H	H	0.875	0.4

Table A1. *Cont.*

File ID	Expert 1	Expert 2	Expert 3	Expert 4	Mean Value of Experts' Judgement	AI Model
8	H	H	M	H	0.6875	0.4
9	H	H	H	H	0.75	0.4
10	H	VH	H	VH	0.875	0.3
11	M	H	H	VH	0.75	0.6
12	M	M	M	M	0.5	0.6
13	M	H	M	H	0.625	0.8
14	L	L	M	L	0.3125	0.4
15	H	H	H	VH	0.8125	0.3
16	H	H	H	H	0.75	0.6
17	L	L	L	M	0.3125	0.3
18	H	H	H	VH	0.8125	0.7
19	H	H	H	H	0.75	0.6
20	VH	VH	H	H	0.875	0.6
21	VH	M	H	H	0.75	0.8
22	L	L	L	L	0.25	0.4
23	VH	H	H	H	0.8125	0.3
24	H	M	M	H	0.625	0.3
25	M	M	H	H	0.625	0.3
26	VH	VH	VH	VH	1	0.8
27	H	H	H	H	0.75	0.3
28	H	H	H	H	0.75	0.3
29	M	H	H	H	0.6875	0.4
30	M	H	M	H	0.625	0.5
31	VH	VH	VH	VH	1	0.9
32	VH	VH	VH	H	0.9375	0.8
33	VH	VH	VH	VH	1	0.8
34	VH	H	H	VH	0.875	0.7
35	L	L	H	M	0.4375	0.3
36	M	M	H	H	0.625	0.6
37	H	M	H	M	0.625	0.7
38	VH	H	VH	VH	0.9375	0.8
39	H	H	VH	H	0.8125	0.4
40	L	M	H	M	0.5	0.4
41	M	H	H	H	0.6875	0.8
42	VH	VH	VH	VH	1	0.9
43	VH	H	H	H	0.8125	0.8
44	M	H	H	H	0.6875	0.3
45	M	M	L	M	0.4375	0.2
46	M	H	H	H	0.6875	0.3
47	H	M	M	M	0.5625	0.4
48	M	M	M	H	0.5625	0.8
49	H	H	M	VH	0.75	0.6
50	M	M	M	H	0.5625	0.7
51	M	H	H	H	0.6875	0.3

Table A2. Rating of real images.

File ID	Expert 1	Expert 2	Expert 3	Expert 4	Mean Value of Experts' Judgement	AI Model
1	L	L	L	M	0.3125	0.3
2	M	M	M	H	0.5625	0.6
3	VH	VH	H	VH	0.9375	0.8
4	H	H	H	H	0.75	0.7
5	L	L	L	M	0.3125	0.2
6	M	M	H	H	0.625	0.4
7	H	M	H	H	0.6875	0.3
8	L	L	L	M	0.3125	0.3
9	H	M	M	H	0.625	0.3
10	M	L	M	M	0.4375	0.3
11	L	M	L	M	0.375	0.3
12	H	M	H	H	0.6875	0.6
13	H	H	H	H	0.75	0.6
14	L	L	L	L	0.25	0.6
15	L	L	L	M	0.3125	0.8
16	L	M	M	L	0.375	0.3
17	H	M	H	M	0.625	0.4
18	H	L	H	H	0.625	0.4
19	H	M	H	H	0.6875	0.7
20	H	H	H	H	0.75	0.8
21	H	H	H	H	0.75	0.7
22	H	H	H	H	0.75	0.7
23	H	H	H	H	0.75	0.4
24	H	H	H	H	0.75	0.7
25	H	M	H	H	0.6875	0.9
26	M	M	M	H	0.5625	0.8
27	VH	M	H	H	0.75	0.7
28	H	M	M	H	0.625	0.3
29	H	H	VH	VH	0.875	0.3
30	M	M	L	M	0.4375	0.6
31	M	M	M	H	0.5625	0.3
32	M	M	M	H	0.5625	0.6
33	H	M	M	H	0.625	0.3
34	H	M	H	H	0.6875	0.4
35	H	H	H	H	0.75	0.7
36	H	M	M	H	0.625	0.3
37	H	H	M	VH	0.75	0.4
38	M	H	M	VH	0.6875	0.7
39	H	H	M	H	0.6875	0.7
40	H	M	M	H	0.625	0.5
41	H	H	H	H	0.75	0.4
42	H	M	M	VH	0.6875	0.3
43	H	M	M	H	0.625	0.7
44	M	H	M	H	0.625	0.6
45	M	H	M	M	0.5625	0.3
46	L	M	H	M	0.5	0.6
47	H	M	H	H	0.6875	0.3
48	M	M	M	H	0.5625	0.7
49	H	M	L	M	0.5	0.6
50	H	M	M	H	0.625	0.8
51	H	M	M	H	0.625	0.3

References

1. ANSA. Available online: https://www.ansa.it/english/news/business/2026/02/03/792-workplace-deaths-in-2025-down-5-from-2024_080391d5-3e67-40e3-a0ea-ca9b9f03350a.html (accessed on 5 February 2026).
2. La Fata, C.M.; Giallanza, A.; Micale, R.; La Scalia, G. Ranking of occupational health and safety risks by a multi-criteria perspective: Inclusion of human factors and application of VIKOR. *Saf. Sci.* **2021**, *138*, 105234. [CrossRef]
3. Uddin, S.M.J.; Albert, A.; Ovid, A.; Alsharef, A. Leveraging ChatGPT to Aid Construction Hazard Recognition and Support. *Saf. Educ. Train. Sustain.* **2023**, *15*, 7121.
4. Eurostat. Accidents at Work Statistics. Available online: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Accidents_at_work_statistics (accessed on 5 February 2026).
5. Jeelani, I.; Albert, A.; Azevedo, R.; Jaselskis, E.J. Development and Testing of a Personalized Hazard-Recognition Training Intervention. *J. Constr. Eng. Manag.* **2017**, *143*, 04016120. [CrossRef]
6. Perlman, A.; Sacks, R.; Barak, R. Hazard Recognition and Risk Perception in Construction. *Saf. Sci.* **2014**, *64*, 22–31. [CrossRef]
7. Sun, J.; Chang, F.; Zhou, Z.; Man, S.-S.; Shou Chan, A.H. A Systematic Review of Hazard Recognition and Risk Perception Research in the Construction Industry. *Saf. Sci.* **2025**, *186*, 106813. [CrossRef]
8. Fleming, M.; Fischer, B. Hazard Recognition: Bridging Knowledge and Competency for Process and Occupational Safety. *Prof. Saf.* **2017**, *62*, 52–61.
9. Hardison, D.; Hallowell, M.; Littlejohn, R. Does the Format of Design Information Affect Hazard Recognition Performance in Construction Hazard Prevention through Design Reviews? *Saf. Sci.* **2020**, *121*, 191–200. [CrossRef]
10. Khaleghi, P.; Akbari, H.; Alavi, N.M.; Kashani, M.M.; Batooli, Z. Identification and Analysis of Human Errors in Emergency Department Nurses Using SHERPA Method. *Int. Emerg. Nurs.* **2022**, *62*, 101159. [CrossRef]
11. Pereira, F.; González García, M.d.l.N.; Poças Martins, J. An Evaluation of the Technologies Used for the Real-Time Monitoring of the Risk of Falling from Height in Construction—Systematic Review. *Buildings* **2024**, *14*, 2879. [CrossRef]
12. Sarvari, H.; Baghbaderani, A.B.; Chan, D.W.M.; Beer, M. Determining the Significant Contributing Factors to the Occurrence of Human Errors in Urban Construction Projects: A Delphi-SWARA Study Approach. *Technol. Forecast. Soc. Change* **2024**, *205*, 123512. [CrossRef]
13. Occupational Safety and Health Administration (OSHA). *OSHA 3071—Job Hazard Analysis*; U.S. Department of Labor: Washington, DC, USA, 2002.
14. Lingard, H.; Rowlinson, S. *Occupational Health and Safety in Construction Project Management*; Routledge: London, UK, 2004.
15. Jeelani, I.; Han, K.; Albert, A. Development of Immersive Personalized Training Environment for Construction Workers. In Proceedings of the Congress on Computing in Civil Engineering, Seattle, WA, USA, 25–27 June 2017; pp. 407–415.
16. La Fata, C.M.; Giallanza, A.; Micale, R.; La Scalia, G. Toward acceptance of human-robot collaboration in industrial settings: A bibliometric and systematic literature review. *Int. J. Adv. Manuf. Technol.* **2025**, *139*, 2139–2160. [CrossRef]
17. Jamwal, A.; Agrawal, R.; Sharma, M.; Giallanza, A. Industry 4.0 Technologies for Manufacturing Sustainability: A Systematic Review and Future Research Directions. *Appl. Sci.* **2021**, *11*, 5725. [CrossRef]
18. Dzeng, R.J.; Lin, C.T.; Fang, Y.C. Using Eye-Tracker to Compare Search Patterns between Experienced and Novice Workers for Site Hazard Identification. *Saf. Sci.* **2016**, *82*, 56–67. [CrossRef]
19. Cheng, B.; Luo, X.; Mei, X.; Chen, H.; Huang, J. A Systematic Review of Eye-Tracking Studies of Construction Safety. *Front. Neurosci.* **2022**, *16*, 891725. [CrossRef]
20. Wang, D.; Chen, J.; Zhao, D.; Dai, F.; Zheng, C.; Wu, X. Monitoring Workers' Attention and Vigilance in Construction Activities through a Wireless and Wearable Electroencephalography System. *Autom. Constr.* **2017**, *82*, 122–137. [CrossRef]
21. Adebayo, Y.; Udoh, P.; Kamudyariwa, X.B.; Osobajo, O.A. Artificial Intelligence in Construction Project Management: A Structured Literature Review of Its Evolution in Application and Future Trends. *Digital* **2025**, *5*, 26. [CrossRef]
22. Trivedi, P.; Alqahtani, F.M. The Advancement of Artificial Intelligence (AI) in Occupational Health and Safety (OHS) across High-Risk Industries. *J. Infrastruct. Policy Dev.* **2024**, *8*, 6889. [CrossRef]
23. Sacks, R.; Perlman, A.; Barak, R. Construction Safety Training Using Immersive Virtual Reality. *Constr. Manag. Econ.* **2013**, *31*, 1005–1017. [CrossRef]
24. Ding, Y.; Ma, J.; Luo, X. Applications of Natural Language Processing in Construction. *Autom. Constr.* **2022**, *136*, 104169. [CrossRef]
25. Sridi, C.; Brigui, S. The Use of ChatGPT in Occupational Medicine: Opportunities and Threats. *Ann. Occup. Environ. Med.* **2023**, *35*, e42. [CrossRef] [PubMed]
26. Kalyan, K.S. A Survey of GPT-3 Family Large Language Models Including ChatGPT and GPT-4. *Nat. Lang. Process. J.* **2024**, *6*, 100048. [CrossRef]
27. GPT-4. Available online: https://openai.com/it-IT/index/gpt-4-research/ (accessed on 13 May 2025).
28. OpenAI. GPT-4 Technical Report. *arXiv* **2023**. [CrossRef]
29. Basulo-Ribeiro, J.; Teixeira, L. Is ChatGPT an Ally or an Enemy? Its Impact on Society Based on a Systematic Literature Review. *J. Inf. Sci. Theory Pract.* **2024**, *12*, 79–95.

30. Saka, A.; Taiwo, R.; Saka, N.; Salami, B.A.; Ajayi, S.; Akande, K.; Kazemi, H. GPT Models in Construction Industry: Opportunities, Limitations, and a Use Case Validation. *Dev. Built Environ.* **2024**, *17*, 100300. [CrossRef]
31. Firat, M. What ChatGPT Means for Universities: Perceptions of Scholars and Students. *J. Appl. Learn. Teach.* **2023**, *6*, 57–63. [CrossRef]
32. Li, J.; Dada, A.; Puladi, B.; Kleesiek, J.; Egger, J. ChatGPT in Healthcare: A Taxonomy and Systematic Review. *Comput. Methods Programs Biomed.* **2024**, *245*, 108013. [CrossRef] [PubMed]
33. Zong, M.; Krishnamachari, B. A Survey on GPT-3. *arXiv* **2022**. [CrossRef]
34. Uddin, S.M.J.; Albert, A.; Tamanna, M. Harnessing the Power of ChatGPT to Promote Construction Hazard Prevention through Design (CHPtD). *Eng. Constr. Archit. Manag.* **2024**, *32*, 7832–7856. [CrossRef]
35. Aladağ, H. Assessing the Accuracy of ChatGPT Use for Risk Management in Construction Projects. *Sustainability* **2023**, *15*, 16071. [CrossRef]
36. Hussain, R.; Sabir, A.; Lee, D.Y.; Zaidi, S.F.A.; Pedro, A.; Abbas, M.S.; Park, C. Conversational AI-Based VR System to Improve Construction Safety Training of Migrant Workers. *Autom. Constr.* **2024**, *160*, 105315. [CrossRef]
37. Xiao, B.; Wang, Y.; Zhang, Y.; Chen, C.; Darko, A. Automated Daily Report Generation from Construction Videos Using ChatGPT and Computer Vision. *Autom. Constr.* **2024**, *168*, 105874. [CrossRef]
38. Samsami, R. Optimizing the Utilization of Generative Artificial Intelligence (AI) in the AEC Industry: ChatGPT Prompt Engineering and Design. *CivilEng* **2024**, *5*, 971–1010. [CrossRef]
39. Bazrafshan, P.; Melag, K.; Ebrahimkhanlou, A. Semantic and lexical analysis of pre-trained vision language artificial intelligence models for automated image descriptions in civil engineering. *AI Civ. Eng.* **2025**, *4*, 17. [CrossRef]
40. Tran, D.Q.; Jeon, Y.; Park, M.; Park, S. GPT-based Logic Reasoning for Hazard Identification in Construction Site using CCTV Data. In Proceedings of the 41st International Symposium on Automation and Robotics in Construction, ISARC 2024, Lille, France, 3–5 June 2024; pp. 291–298.
41. Barone, G. *Machine Learning e Intelligenza Artificiale: Metodologie per lo Sviluppo di Sistemi Automatici*; Dario Flaccovio Editore: Palermo, Italy, 2021.
42. European Union. *Regulation (EU) 2024/1689*; Official Journal of the European Union: Luxembourg, 2024; L 2024/1689.