How Reliable Are Large Language Models? Zero-Shot Detection of Construction Hazards

Nishi Chaudhary[1], S M Jamil Uddin[1], Mahzabin Tamanna[2], Alex Albert[3], Abdur Rahman Bin Shahid[4]

[1]Department of Construction Management, Colorado State University, USA
[2]Department of Computer Science, North Carolina State University, USA
[3]Department of Civil, Construction, and Environmental Engineering, North Carolina State University, USA
[4]School of Computing, Southern Illinois University, USA
nishi.chaudhary@colostate.edu, smj.uddin@colostate.edu, mtamann@ncsu.edu, alex_albert@ncsu.edu, shahid@cs.siu.edu

Abstract – The construction industry persistently underperforms in hazard recognition, often leading to severe workplace injuries due to unrecognized hazards. With the recent advancements in Artificial Intelligence (AI) and the emergence of Large Language Models (LLM), the construction sector has begun exploring these technologies for various applications. However, a systematic comparison of popular LLMs to evaluate their effectiveness in identifying construction hazards remains unexplored. Additionally, previous studies have primarily focused on assessing LLMs using textual input and output, leaving their performance with visual inputs underexplored. This study addresses this gap by systematically assessing and comparing the hazard recognition performance of five widely used LLMs using construction case images. The findings establish a baseline standard for LLMs in construction hazard identification through zero-shot learning and reveal that LLMs do not perform significantly well in this context. Additionally, the study provides valuable insights into the reliability and potential applications of LLMs for enhancing hazard recognition in the construction industry.

Keywords – LLM, Hazard Recognition, Construction Safety, Image Analysis

1 Introduction

The construction industry plays a pivotal role in developing and maintaining the infrastructure, thereby shaping broader society [1]. Despite its importance, the construction industry reports a disproportionate number of safety incidents around the world [2]. These incidents often result in fatal and non-fatal injuries to the workers. In the United States alone, approximately 1,000 fatal injuries take place every year and the number go well beyond 200,000 for non-fatal injuries [3]. In addition to the fatal and non-fatal injuries, these incidents also cause a significant amount of economic damage [4].

Over the years, researchers and industry professionals have put much effort into investigating the reasons behind these safety incidents. Evidence suggests that one of the main reasons for these incidents is the failure to identify construction hazards [5, 6]. Studies show that the construction industry performs poorly in recognizing workplace hazards, with over 70% of work-related incidents attributed to poor hazard recognition [7].

Several tools have been developed over the years to improve the hazards recognition efforts in the construction industry. For example, job hazard analysis (JHA) tool is used to catalog hazards that are associated with specific construction tasks. However, this method assumes that workers possess a level of proficiency in identifying hazards, which in reality is not the case [8, 9]. On the other hand, tools such as safety checklists fail to contribute due to their limitation of identifying a limited number of hazards [10]. A number of studies have demonstrated that, despite the usage of such safety tools, over 40% of construction hazards remain unrecognized in the construction workplaces and they in turn cause these fatal and non-fatal incidents [11, 12].

On the technological front, several other tools have been developed to aid hazard recognition over the years. For example, usage of Building Information Modeling (BIM) has been popular among researchers and industry professionals [13]. BIM enables collaboration, visualization, and enhances the hazard identification process. Other efforts have focused on technologies such as eye tracking, augmented reality (AR), and virtual reality (VR) environments [14, 15]. While these technological solutions aid the hazard recognition process, they have their own limitations. For instance, these technology-driven tools require significant expertise, development time, resources, and financial investments. These barriers often hinder widespread

adoption of these technological solutions.

In recent years with the development and uprising of Artificial Intelligence (AI) and Large Language Models (LLM), the construction industry has also leaned towards adopting these solutions to improve the safety situation of construction workplaces. Different studies have demonstrated how some popular LLMs, such as ChatGPT, can be beneficial in improving construction safety [16-19]. In light of recent advancements, this study focuses on evaluating the capabilities of five different LLMs in identifying construction hazards using zero-shot learning. This evaluation aims to establish a baseline standard for LLM performance with no prior training, additional knowledge base, or guidelines, and offer insights on the reliability of LLMs in recognizing construction hazards.

2 Use of LLM in Construction Safety

In recent years, Large Language Models (LLM) have gained significant popularity across different industries and domains. Construction industry is no different from others. Although the industry is in its very early stages of adopting LLMs on a full scale for diverse applications, several studies have explored the opportunities these LLMs present. For example, a number of studies explored the possibility of integrating LLM and BIM to support information retrieval from the building models [19-21]. Other studies have focused on leveraging different LLMs for project management tasks such as automated sequence planning [22], generating construction schedules [23], automated classification of contractual risk clauses [24], automatic matching of look ahead planning tasks [25] etc. Additionally, some studies focused on how to effectively use LLMs to improve the construction education outcomes [16, 26, 27].

On the safety front, several studies have examined the usability of LLMs to improve the health and safety condition of construction workplaces. For example, Uddin et al. [18] conducted a controlled experiment with 42 construction engineering students. Their effort demonstrated that LLMs can be particularly effective in aiding hazard recognition efforts. They also suggest that LLMs can be integrated as part of safety education for construction students albeit with caution. Another study by Uddin et al. [17] explored the usability of LLM in aiding construction hazard prevention through design efforts. The study demonstrated that LLM can improve the hazard recognition efforts during the design phase by approximately 40%. Wang et al. [28] evaluated LLM's ability to extract causal factors from construction accident reports. The study found that LLM can perform well as an assisting tool, offering clear and reliable insights, but it still requires further development for professional applications like crane safety. The research highlights the potential of LLMs in construction while emphasizing the need for refinement to enhance their practical utility.

Smetana et al. [29] leveraged an LLM model to analyze textual data from OSHA's Severe Injury Reports (SIR) for highway construction accidents. Using advanced NLP techniques, clustering, and LLM prompting, they identified major accident types, including heat-related and struck-by injuries, while uncovering commonalities between incidents. The findings demonstrate the potential of AI and LLMs to enhance data-driven safety analysis and support the development of more effective prevention strategies in the highway construction industry. Hussain et al. [30] developed a virtual reality-based safety training system incorporating LLM as a live AI instructor to address communication barriers and trainer limitations, particularly for migrant workers. Testing across five countries showed a 23% improvement in knowledge scores, demonstrating the system's effectiveness. The research highlights the system's potential to improve safety training globally, reduce construction site accidents, and advance immersive and AI-driven training methodologies.

While the existing studies have demonstrated the potential of LLMs in various aspects of construction including workers' health and safety, they predominantly focused on evaluating a single LLM and relied exclusively on textual inputs and outputs. None of these efforts explored a comparative analysis of multiple LLMs to assess their relative effectiveness in achieving safety outcomes. Additionally, the studies did not investigate the use of image-based inputs for hazard recognition, despite the growing capabilities of modern LLMs to process and interpret visual data.

With recent advancements in LLMs enabling them to analyze multimodal inputs, including images, it becomes crucial to evaluate whether these models can effectively extract safety-critical information from construction images and identify hazards. Such an investigation would address a significant gap in the literature and provide insights into the broader applicability of LLMs for improving construction safety practices.

Hence, this study focuses on assessing five different LLMs' capability of identifying potential hazards from construction case images and then conducting a comparative analysis to demonstrate if all the LLMs are reliable and if they perform equally in identifying potential hazards from construction case images.

3 Methodology

To achieve the goals of this study, it was necessary to establish a ground truth for evaluating the hazard recognition capabilities of the LLMs. This was

accomplished by using a dataset of 16 different images of construction activities captured as part of a previous study [31]. These images depict a wide variety of construction activities such as welding, cutting, drilling, crane rigging among others.

Once gathered, these images were then analyzed by a panel of 17 construction industry safety experts. The collective experience of this panel was over 300 years at the time of this assessment. The expert panel was tasked with analyzing the images and identifying all the potential safety hazards for each image. One example image with annotated safety hazards is shown in Figure 1. The expert panel identified a total of 120 construction hazards from these 16 case images. These pre-identified hazards are going to serve as ground truth for our LLM assessment experiment.

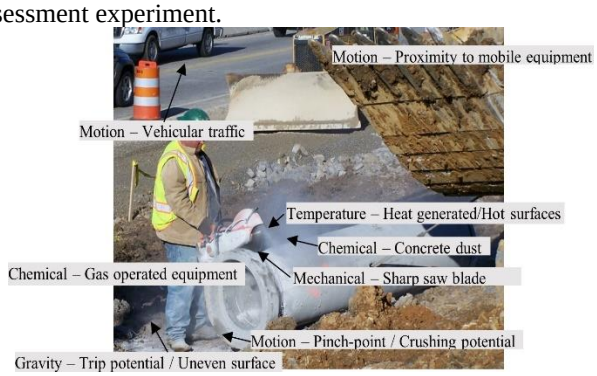

Figure 1. Example image of construction activity with pre-identified hazards.

Next, we chose five LLMS, i.e., ChatGPT 4, GPT 4o, GPT o1, Gemini 2.0, and Claude for this experiment. The selection of these five LLMs was based on their industry adoption [17, 32, 33], multimodal capabilities [34, 35], architectural diversity [36], and expected performance variation. These models are widely used and include multimodal processing which is crucial for hazard recognition from construction case images. The study specifically leveraged the zero-shot learning capabilities of these models, allowing them to identify hazards without requiring domain-specific training or environment customization. This approach was deemed particularly relevant to assess the baseline performance of these LLMs without any training or additional knowledge input.

To ensure consistency and standardization, a systematic approach was developed to input all the images into these LLMs for hazard identification. First, the following prompt was provided:

"You're a construction safety expert. You will be provided with 16 different construction images. You will have to analyze these images and identify the potential safety hazards for each image."

Then each image was fed into to LLM chat prompt along with the following instruction:

"Identify all potential safety hazards for the construction activity in this image."

No additional contextual or domain-specific prompts were provided, ensuring the zero-shot learning approach remained intact. This standardized methodology facilitated comparability of outputs across all the models.

Once the LLMs returned their results for each image, the hazards were then recorded carefully by the research team ensuring traceability and consistency for subsequent analysis. The process was conducted under controlled experimental conditions, ensuring that all models were evaluated using identical inputs and queries.

The output generated by the LLMs were then compared against the ground truth dataset to evaluate their hazard recognition performance. Three different metrics were used to evaluate the performance of these LLMs, i.e., Precision, Recall, and F1-score [37, 38]. Precision score denotes the proportion of correctly identified hazards relative to the total hazards identified by the model. Recall is the proportion of correctly identified hazards relative to the total hazards present in the ground truth dataset. And F1-score is the harmonic mean of precision and recall, providing a balanced measure of accuracy and completeness of the models.

4 Data Analysis and Results

In order to measure the Precision, Recall, and F1-Score of each LLM, we used equations 1, 2, and 3 respectively for each image.

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \quad (1)$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \quad (2)$$

$$F1 - Score_i = 2 \times \frac{Precision_i \times Recall_i}{Precision_i + Recall_i} \quad (3)$$

Where:

TP_i = Hazards correctly identified by each LLM
FP_i = Hazards incorrectly identified by each LLM
FN_i = Hazards that exist in the ground truth but were missed by each LLM

Once the Precision, Recall, and F1-Score for all 16 images were gathered, then they were averaged to get the mean Precision, Recall, and F1-Score using equations 4, 5, and 6 where n is equal to 16.

$$Avg.Precision = \frac{1}{n}\sum_{i=1}^{n} Precision_i \quad (4)$$

$$Avg.Recall = \frac{1}{n}\sum_{i=1}^{n} Recall_i \quad (5)$$

$$Avg.F1-Score = \frac{1}{n}\sum_{i=1}^{n} F1-Score_i \quad (6)$$

Table 1 below shows the summarized results for the five LLM models that were assessed.

Table 1 Summary of LLM Performance

Models	Precision	Recall	F1-Score
ChatGPT-4	0.30	0.29	0.30
GPT4o	0.25	0.30	0.27
GPTo1	0.27	0.30	0.27
Gemini 2.0	0.31	0.33	0.31
Claude	0.42	0.29	0.34
Overall	0.31	0.30	0.30

4.1 ChatGPT-4

As can be seen in Table 1, ChatGPT-4 achieved 0.30, 0.29, and 0.30 scores in Precision, Recall, and F1-Score metrics respectively. A precision score of 0.30 indicates that only 30% of the hazards identified by the model were accurate, which also highlights a significant proportion of false positives. This indicates that ChatGPT-4's level of precision in recognizing construction hazards as a standalone tool may not be very reliable. Similarly, the recall score of 0.29 demonstrates the model's inability to detect the majority of the hazards present in the actual scenarios. The F1-score of 0.30 shows that the model struggles to achieve both accuracy and completeness in construction hazard identification from images.

4.2 GPT4o

GPT4o's performance in construction hazard recognition from the images is similar to GPT4's. GPT4o's precision score of 0.25 indicates that only 25% of the identified hazards were accurate, reflecting a higher rate of false positives. Additionally, the recall score of 0.30 shows that the model was able to identify only 30% of the actual hazards in the test scenario. The F1-score of 0.27 indicates that this model also struggles to identify construction hazards on its own from the construction images.

4.3 GPTo1

GPTo1 model performed rather poorly in construction hazard identification in this study. GPTo1 achieved 0.27 in precision which demonstrates a poor performance in recognizing hazards accurately, with a higher level of false positives. The recall score of 0.30, which is consistent with GPT4 and GPT4o, shows that the model was able to identify 30% of the actual hazards present in the test scenarios. This underscores that a significant proportion of hazards went unrecognized. The F1-score of 0.27, which is similar to GPT4o, highlights the model's limitations in recognizing construction hazards successfully.

4.4 Gemini 2.0

Gemini 2.0 with precision score of 0.31, recall score of 0.33, and F1-score of 0.31 demonstrated a slight improvement over the other three LLMs. The precision score indicates that 31% of the hazards were identified accurately. While the recall score of 0.33 demonstrates a higher percentage of correctly identified hazards from the hazards present in the ground truth dataset. The F1-score of 0.33 shows the balanced performance between precision and recall for this model.

4.5 Claude

Claude's performance in construction hazard recognition with a precision score of 0.42 demonstrates its ability to identify hazards with relatively higher accuracy, which is the highest among other models tested in this study. This score also highlights a significantly lower rate of false positives. However, a lower recall score of 0.29 indicates that Claude identified only 29% of the actual hazards from the ground truth dataset. While its precision ensures that the hazards identified are likely correct, the lower recall suggests that many actual hazards remain unrecognized, which could be a concern in environments requiring exhaustive risk identification.

4.6 Trend of Hazard Identification

Along with assessing different LLMs' hazard recognition performance, it is also important to examine if there's a common trend demonstrated in construction hazard recognition. To achieve this, we aggregated all the hazards that were correctly identified by the LLMs, referred to as True Positives (TP), to gain insight into which hazards were most successfully recognized. We also summed up all the False Negatives (FN) to identify the hazards that were least frequently recognized by the LLMs. Figure 2 and figure 3 below show the top five and bottom five hazards identified by the LLMs.

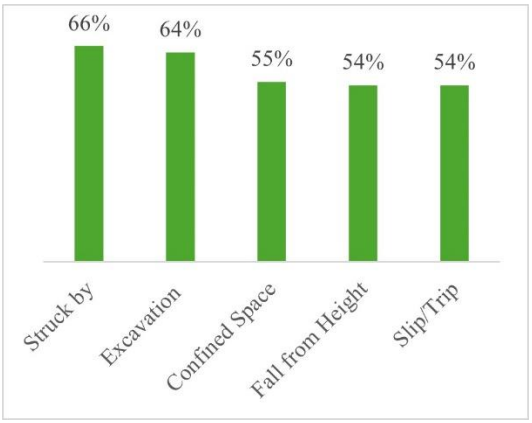

Figure 2. Top five hazards identified by the LLMs.

As can be seen in figure 2, the LLMs overall performed better in identifying more common hazards such as struck by, excavation, confined space, and fall hazards.

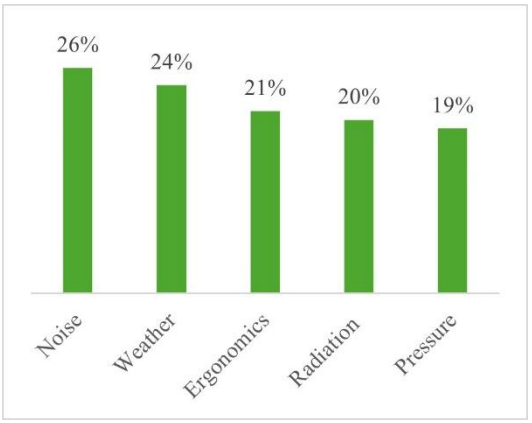

Figure 3. Least five hazards identified by LLMs.

On the other hand, figure 3 demonstrates that the LLMs performed rather poorly in identifying hazards such as noise, weather, ergonomics, radiation, and pressure. This finding is consistent with previous research efforts where the studies evaluated the hazard recognition trend of the construction workforce [12, 39].

5 Discussion

The performance of the evaluated LLMs in construction hazard recognition from images under zero-shot learning conditions reveals a varied outcome across all the measured metrics. Among the five tested models, Claude showed the highest precision level in identifying construction hazards. This suggests that Claude minimizes false positives better than the other models, making it particularly suitable for applications where avoiding unnecessary interventions is critical. However, the lower recall value of Claude indicates the failure to identify all the hazards present in the ground truth dataset. On the other hand, Gemini 2.0 achieved the highest recall score and maintained a balanced F1-score. Compared to Claude, Gemini 2.0 sacrifices some precision but compensates with better recall, making it a more reliable choice for scenarios requiring broader hazard identification.

The superior performance of Claude in precision may be attributed to its underlying architecture, which possibly favors conservative predictions by avoiding overgeneralizations. Meanwhile, Gemini 2.0's balanced performance could stem from a more robust generalization capability, enabling it to identify a larger proportion of hazards without excessively sacrificing precision.

The remaining three models, ChatGPT-4, GPT4o, and GPT-o1 exhibited comparable performance, with precision ranging from 0.25 to 0.30 and recall clustering around 0.30. These models struggled to achieve a balance between precision and recall, with their F1 scores hovering at 0.27 to 0.30. Notably, ChatGPT-4 demonstrated the highest F1 score among this group at 0.30, indicating a slightly better overall performance in balancing accuracy and coverage. The relatively lower precision of GPT4o (0.25) suggests a greater tendency toward false positives, while GPT-o1 showed a slight improvement in reducing these errors.

The overall average metrics across all models, precision of 0.31, recall of 0.30, and F1 score of 0.30, highlight a consistent challenge for LLMs in performing hazard recognition tasks as standalone tools under zero-shot learning conditions. However, it is worth noting that the hazard recognition performance of LLMs using zero-shot learning is comparable to that of human participants with no prior intervention or training, as observed in previous studies [12, 31, 40].

The variation in precision, recall, and F1-score across LLMs can be attributed to differences in model architecture, training data, and inference strategies. For example, Claude exhibits high precision but low recall as it applies conservative thresholds, flagging only high-confidence hazards to minimize false positives. In contrast, Gemini 2.0, designed as a multimodal AI, prioritizes higher recall over precision, leading to more hazard detections but also increased false positives. The GPT models from OpenAI are balanced models that achieve moderate precision and recall, making them consistent but not specialized in hazard detection.

The results obtained from this experiment indicate that LLMs face significant challenges in construction hazard recognition from images when they are used without domain specific knowledge, fine tuning, or specific guidance. The findings contrast with previous studies that showed LLMs perform exceedingly well when they are integrated into safety training or education contexts for construction professionals. Previous studies

used LLMs to support humans by generating insights or suggesting hazards, leveraging human reasoning to fill gaps in understanding [17, 18]. While on the other hand, the current zero-shot learning setup relied entirely on the models' ability to independently interpret the construction images and identify hazards. This exposes the limitations of the LLMs' specific knowledge base.

These findings of this study demonstrate that, while LLMs perform remarkably well with structured prompts, and additional knowledge base, without these interventions and without human assistance, LLM's can perform rather poorly in construction hazard recognition.

6 Conclusion

Construction hazard recognition is one of the pivotal steps in order to ensure a safer construction environment. Previous studies have demonstrated that construction hazards often remain unrecognized in the workplace which translates to both fatal and non-fatal accidents. In order to tackle this issue, numerous studies have focused on developing and adopting various tools to help with hazard recognition.

With the advancement of modern technology, Artificial Intelligence (AI), and Large Language Models (LLM), researchers and practitioners have started to utilize these technological solutions to improve construction hazard recognition. However, since there are multiple different LLM platforms available, it was necessary to conduct a comparative analysis to see if they are reliable and which model performs better than others. Additionally, it was also important to assess these models' performance with zero-shot learning, in other words, with no enforced knowledge.

This study was designed carefully to evaluate five different LLMs performance in construction hazard recognition from construction case images. The findings of this study reveal that LLMs perform rather poorly in identifying construction hazards with zero-shot learning as standalone tool. However, the hazard recognition trend demonstrates that the LLMs' performances are not too different from the human subjects from construction industry. LLMs show a similar trend in identifying common hazards and missing out uncommon hazards, similar to previous other studies [12, 39, 41].

The study highlights that although LLMs show a promise in identifying construction hazards, they may require some input, training, additional knowledge base, interventions, and human assistance to perform better as previous studies have demonstrated an improved hazard recognition performance by human subjects with assistance from different LLMs [17, 18].

In future work, we plan to explore the impact of fine-tuning LLMs using larger construction-specific datasets to assess their potential for improved hazard recognition. While our current study leveraged zero-shot learning to establish baseline performance, real-world applications would likely require models that have been trained on domain-specific data. We plan to expand our analysis to include a detailed examination of false positives and false negatives for each model, providing deeper insight into the specific types of hazards that LLMs fail to recognize or misclassify. This will help identify recurring patterns in model errors and guide improvements in hazard recognition accuracy through targeted fine-tuning and dataset enhancements.

Additionally, building on our analysis of hazard identification trends in Section 4.6, future work will incorporate qualitative insights from construction safety experts to better understand the nature of LLM errors. While our study quantitatively identified the types of hazards frequently missed by LLMs, expert consultation will provide deeper context on why certain hazards are overlooked and how models can be improved. Specifically, we will work with industry professionals, site safety managers, and researchers to identify common misclassifications, potential biases in AI hazard detection, and effective strategies for model training.

References

[1] Labor, U.S.D.o., *Civil Engineers : Occupational Outlook Handbook: : U.S. Bureau of Labor Statistics*, in *2021*. 2021.

[2] Statistics, U.S.B.o.L., *Construction and Extraction Occupations : Occupational Outlook Handbook: : U.S. Bureau of Labor Statistics*. 2021.

[3] Bls, *Injuries, Illnesses, and Fatalities*, in *U.S. Bureau of labor Statistics*. 2021.

[4] Waehrer, G.M., et al., *Costs of occupational injuries in construction in the United States.* Accident Analysis and Prevention, 2007. **39**(6): p. 1258-1266.

[5] Perlman, A., R. Sacks, and R. Barak, *Hazard recognition and risk perception in construction.* Safety Science, 2014. **64**: p. 22-31.

[6] Carter, G. and S.D. Smith, *Safety Hazard Identification on Construction Projects.* JOURNAL OF CONSTRUCTION ENGINEERING AND MANAGEMENT, 2006. **132**(2): p. 197-205.

[7] Haslam, R.A., et al., *Contributing factors in construction accidents.* Applied Ergonomics, 2005. **36**: p. 401-415.

[8] Jeelani, I., A. Albert, and J.A. Gambatese, *Why Do Construction Hazards Remain Unrecognized at the Work Interface?* Journal of Construction Engineering and Management, 2017. **143**(5): p. 1-10.

[9] Rozenfeld, O., et al., *Construction Job Safety Analysis.* Safety Science, 2010. **48**(4): p. 491-498.
[10] Guo, B.H., T.W. Yiu, and V.A. González, *Predicting safety behavior in the construction industry: Development and test of an integrative model.* Safety science, 2016. **84**: p. 1-11.
[11] Jeelani, I., et al., *Development and Testing of a Personalized Hazard-Recognition Training Intervention.* Journal of Construction Engineering and Management, 2017. **143**(5): p. 04016120-04016120.
[12] Uddin, S.M.J., et al., *Hazard Recognition Patterns Demonstrated by Construction Workers.* International Journal of Environmental Research and Public Health, 2020. **17**(21): p. 7788-7788.
[13] Kim, I., Y. Lee, and J. Choi, *BIM-based hazard recognition and evaluation methodology for automating construction site risk assessment.* Applied Sciences (Switzerland), 2020. **10**(7).
[14] Jeelani, I., et al., *Are Visual Search Patterns Predictive of Hazard Recognition Performance? Empirical Investigation Using Eye-Tracking Technology.* Journal of Construction Engineering and Management, 2019. **145**(1): p. 04018115-04018115.
[15] Li, X., et al., *A critical review of virtual and augmented reality (VR/AR) applications in construction safety.* Automation in Construction, 2018. **86**: p. 150-162.
[16] Uddin, S.M.J., et al., *ChatGPT as an Educational Resource for Civil Engineering Students.* Computer Applications in Engineering Education, 2024.
[17] Uddin, S.M.J., A. Albert, and M. Tamanna, *Harnessing the power of ChatGPT to promote Construction Hazard Prevention through Design (CHPtD).* Engineering, Construction and Architectural Management, 2024.
[18] Uddin, S.M.J., et al., *Leveraging ChatGPT to Aid Construction Hazard Recognition and Support Safety Education and Training.* Sustainability, 2023. **15**(9): p. 7121-7121.
[19] Zheng, J. and M. Fischer, *BIM-GPT: a Prompt-Based Virtual Assistant Framework for BIM Information Retrieval.* 2023/04/18.
[20] Rane, N., S. Choudhary, and J. Rane, *Integrating Building Information Modelling (BIM) with ChatGPT, Bard, and similar generative artificial intelligence in the architecture, engineering, and construction industry: applications, a novel framework, challenges, and future scope.* Bard, and similar generative artificial intelligence in the architecture, engineering, and construction industry: applications, a novel framework, challenges, and future scope (November 22, 2023), 2023.
[21] Zheng, J. and M. Fischer, *Dynamic prompt-based virtual assistant framework for BIM information search.* Automation in Construction, 2023. **155**: p. 105067.
[22] You, H., et al., *Robot-enabled construction assembly with automated sequence planning based on ChatGPT: RoboGPT.* Buildings, 2023. **13**(7): p. 1772.
[23] Prieto, S.A., E.T. Mengiste, and B. García de Soto, *Investigating the Use of ChatGPT for the Scheduling of Construction Projects.* Buildings, 2023. **13**(4): p. 857-857.
[24] Moon, S., S. Chi, and S.-B. Im, *Automated detection of contractual risk clauses from construction specifications using bidirectional encoder representations from transformers (BERT).* Automation in Construction, 2022. **142**: p. 104465.
[25] Amer, F., Y. Jung, and M. Golparvar-Fard, *Transformer machine learning language model for auto-alignment of long-term and short-term plans in construction.* Automation in Construction, 2021. **132**: p. 103929.
[26] Abril, D.E., M.A. Guerra, and S.D. Ballen. *ChatGPT to Support Critical Thinking in Construction-Management Students*. in *2024 ASEE Annual Conference & Exposition*. 2024.
[27] Zhao, T., et al., *Impact of ChatGPT on Student Writing in Construction Management: Analyzing Literature and Countermeasures for Writing Intensive Courses.* Proceedings of 60th Annual Associated Schools, 2024. **5**: p. 339-348.
[28] Wang, Y., et al. *Investigating the Potential of ChatGPT in Construction Management: A Study of Interpreting Construction Crane-Related Accident Reports*. in *International Symposium on Advancement of Construction Management and Real Estate*. 2023. Springer.
[29] Smetana, M., et al., *Highway Construction Safety Analysis Using Large Language Models.* Applied Sciences, 2024. **14**(4): p. 1352.
[30] Hussain, R., et al., *Conversational AI-based VR system to improve construction safety training of migrant workers.* Automation in Construction, 2024. **160**: p. 105315.
[31] Albert, A., M.R. Hallowell, and B.M. Kleiner, *Enhancing Construction Hazard Recognition and Communication with Energy-Based Cognitive Mnemonics and Safety Meeting Maturity Model: Multiple Baseline Study.* Journal of Construction Engineering and Management, 2014. **140**(2): p. 04013042-04013042.
[32] Genc, O. and O. Genc, *Assessing the role of AI in advancing construction sector industrial symbiosis research: a comparative study of leading digital*

assistants. Environment, Development and Sustainability 2024, 2024-12-04.
[33] Ahmadi, E., S. Muley, and C. Wang, *Automatic construction accident report analysis using large language models (LLMs).* Journal of Intelligent Construction, 2025/1/1. **3**(1).
[34] OpenAI. *ChatGPT can now see, hear, and speak*. 2023 [cited 2025 3/12/2025]; Available from: https://openai.com/index/chatgpt-can-now-see-hear-and-speak/.
[35] Claude. *Introducing the next generation of Claude*. 2024 [cited 2025 3/12/2025]; Available from: https://www.anthropic.com/news/claude-3-family.
[36] Naveed, H., et al., *A comprehensive overview of large language models.* arXiv preprint arXiv:2307.06435, 2023.
[37] Yacouby, R. and D. Axman, *Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models.* Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems, 2020/11.
[38] Orozco-Arias, S., et al., *Measuring Performance Metrics of Machine Learning Algorithms for Detecting and Classifying Transposable Elements.* Processes 2020, Vol. 8, Page 638, 2020-05-27. **8**(6).
[39] Albert, A., B. Pandit, and Y. Patil, *Focus on the fatal-four: Implications for construction hazard recognition.* Safety Science, 2020. **128**(August 2019): p. 104774-104774.
[40] Albert, A., M.R. Hallowell, and B.M. Kleiner, *Enhancing construction hazard recognition and communication with energy-based cognitive mnemonics and safety meeting maturity model: Multiple baseline study.* Journal of Construction Engineering and Management, 2014. **140**(2).
[41] Albert, A., et al., *Empirical measurement and improvement of hazard recognition skill.* Safety Science, 2017. **93**: p. 1-8.