CHATGPT AS A DIGITAL SAFETY ADVISOR: ASSESSING THE EFFICIENCY OF CHATGPT IN PRODUCING SITE-SPECIFIC SAFETY GUIDELINES

Dip Kumar Kundu*[1],Md. Kawsar Akon[2], Sheikh Azim Ur Rashid[3], Ismat Jahan[4] ,Susmita Chowdhury Raka[5] , Md. Rakibul IslamChowdhury[6]

[1]*Department of Building Engineering and Construction Management, KUET, Bangladesh, email:dipk975@gmail.com*
[2]*Department of Building Engineering and Construction Management, KUET, Bangladesh, email:Mdkawsarakon299@gmail.com*
[3]*Assistant Engineer, Local Government Engineering Department, Bangladesh, email:azim.lged@gmail.com*
[4]*Department of Building Engineering and Construction Management, KUET, Bangladesh, email: ismatbd26@gmail.com*
[5]*Department of Building Engineering and Construction Management, KUET, Bangladesh, email: susmitaraka@gmail.com*
[6]*Department of Building, Civil and Environmental Engineering, Concordia University, Montreal, Canada email:crakib2@gmail.com*

***Corresponding Author**

ABSTRACT

Effective construction safety practice depends on accurate hazard identification and on translating regulatory requirements into actionable, site-specific controls. In Bangladesh, the Safety and Rights Society (SRS) reported that construction accounts for approximately 12.1% of workplace fatalities, underscoring the need for scalable decision-support tools for dynamic jobsites. This study evaluates ChatGPT as a language-model-based safety advisor for fall protection by examining (i) the accuracy of hazard detection from textual site scenarios, (ii) the alignment of recommended controls with OSHA 29 CFR 1926 Subpart M and BNBC 2020 requirements, and (iii) expert judgments of clarity, completeness, and feasibility. A scenario-based case study representing working-at-height conditions on a multi-storey building project was developed and used to generate structured ChatGPT outputs, including hazard lists, control measures, and safety checklists. The responses were benchmarked against regulatory provisions and reviewed by experienced construction safety professionals using quantitative scoring and qualitative ratings. Results indicate that ChatGPT can identify close to 90% of hazards in the evaluated scenarios and shows strong alignment with common OSHA fall-protection controls, particularly for standardized measures such as guardrails and personal fall arrest systems. Performance decreases in complex or multi-hazard situations and in requirements involving numeric thresholds, conditional rules, and BNBC-specific procedural details, indicating that competent-person oversight remains essential for field deployment. Overall, the findings suggest that ChatGPT can serve as a scalable safety co-pilot for early-stage planning, documentation, and training support, while highlighting safeguards needed to ensure reliable compliance-oriented implementation.

Keywords: *Large language model, ChatGPT, Fall hazards, Digital safety advisor, Safety Standards (OSHA & BNBC).*

1. INTRODUCTION

Construction remains one of the most hazardous industries globally, exposing workers to a variety of dynamic risks including falls from height, being struck by equipment, caught-in or between incidents, and exposure to hazardous materials. According to recent statistics, the construction sector accounts for approximately 12.1% of global occupational fatalities, with falls representing the leading cause of death (OSHA, 2020a). This underlines the persistent challenge of managing safety in complex, evolving work environments. Despite decades of regulatory oversight and safety management practices, the prevalence of accidents indicates that traditional measures alone are insufficient to address real-time, site-specific hazards.

In the United States, the Occupational Safety and Health Administration (OSHA**)** issues the *29 CFR 1926 Subpart M* standard, which mandates fall protection measures, scaffolding safety requirements, personal protective equipment (PPE) usage, and hazard identification procedures (OSHA, 2020b). Similarly, the Bangladesh National Building Code (BNBC 2020) specifies safety requirements for construction activities, including scaffolding, temporary supports, safe access, and worker training (BNBC, 2020). These standards are comprehensive but generalized and static, often missing dynamic site complexities such as temporary openings, irregular layouts, or changing material storage. Translating them into actionable, scenario-specific guidance remains challenging for safety managers (Mohy et al., 2025; OSHA, 2021).

Recent advances in Artificial Intelligence (AI), particularly Natural Language Processing (NLP), offer an opportunity to bridge the gap between static regulations and dynamic site contexts. Recent advances in Artificial Intelligence (AI), particularly Natural Language Processing (NLP), offer an opportunity to bridge the gap between static regulations and dynamic site contexts. AI-driven advisory systems can support hazard recognition and decision-making by converting textual descriptions of site conditions into structured, context-aware safety recommendations (Ding, Ma, & Luo, 2022; Tixier et al., 2016; Vallabh et al., 2016). Large language models (LLMs) such as ChatGPT are especially relevant because they can interpret natural-language site narratives and generate guidance that can be aligned with regulatory language and safety best practices (Mohy et al., 2025). AI has been explored for construction safety using sensing, computer vision, and analytics. However, rigorous evaluation of LLM-based safety advisors remains limited, especially for generating regulation-consistent, scenario-specific guidance and accurately identifying hazards from text (Imran, Shakir, & Qureshi, 2022; Xie et al., 2023).

This study evaluates ChatGPT as a digital advisor for fall protection by examining (1) the extent to which its recommendations align with OSHA and BNBC requirements, (2) its ability to identify hazards and missing controls of site scenarios, and (3) expert assessments of the accuracy, completeness, and usability of the generated guidance. By systematically testing ChatGPT across representative construction scenarios, this work aims to clarify where LLM-generated safety guidance is reliable, where it fails, and what safeguards are necessary for practical deployment.

This study is significant in advancing evidence-based approaches to AI-assisted construction safety management. By systematically assessing an LLM's ability to interpret OSHA and BNBC provisions and generate implementable, site-specific fall-protection guidance, the research provides a rigorous basis for designing scalable decision-support systems suitable for dynamic jobsite environments. Ultimately, this work provides a foundation for future AI applications in construction safety, offering a scalable, context-aware, and regulation-compliant approach to protecting workers in complex and dynamic environments.

2. LITERATURE REVIEW

2.1 Construction Safety Guidelines

Construction safety is primarily regulated by OSHA's *29 CFR 1926* standards and the Bangladesh National Building Code (BNBC 2020), which outline requirements for fall protection, scaffolding, machinery operations, PPE, and hazard control (Institute, 2020; OSHA, 2020a; OSHA, 2020b; SRS, 2024). These regulations offer consistent technical thresholds, such as guardrail strength and Personal fall arrest systems (PFAS) anchorage capacity, ensuring uniform safety practices across projects (OSHA 2020a; OSHA, 2020b). Despite their strengths, OSHA and BNBC provide generalized safety rules that may not fully address dynamic, project-specific site conditions (OSHA, 2020a; OSHA, 2020b). Research shows that static regulations often fail to capture complex hazard interactions in evolving construction environments (Davis & Conlon, 2017). This indicates the need for adaptive, context-specific safety-support mechanisms.

2.2 Artificial Intelligence in Construction Safety

AI applications in construction safety typically fall into three key categories. First, computer vision–based monitoring detects unsafe behaviors and conditions, including missing Personal Protective Equipmenr (PPE), unsafe postures, and fall-risk zones (Latosinski et al., 2020; Mohy et al., 2025). Second, predictive modeling uses machine learning to estimate incident likelihood and identify leading risk indicators from historical or operational data (H. Assaad et al., 2020). Third, IoT- and sensor-based systems support real-time risk assessment using wearables and on-site monitoring networks (Wolfartsberger, 2019). Although these technologies significantly enhance monitoring capabilities, most existing AI systems focus on *detection only*. They cannot interpret detected hazards within the framework of specific OSHA or BNBC clauses, leaving a gap in generating prescriptive, regulation-aligned safety guidance.

2.3 NLP and ChatGPT Applications

Natural Language Processing (NLP) offers a complementary direction by enabling systems to interpret regulatory documents, safety procedures, and textual descriptions of site conditions. Large language models (LLMs), including ChatGPT, can generate structured outputs such as checklists, hazard explanations, and recommended controls, potentially supporting faster safety planning and communication (Chung et al., 2023; Devlin et al., 2018). In parallel, construction safety management often depends on lessons learned and the reuse of prior project knowledge, which increases the value of tools that can organize and retrieve large volumes of unstructured construction industry data (Azhar, 2011).

Recent generative AI tools (e.g., ChatGPT, Gemini, Copilot) leverage NLP to produce human-readable recommendations and summaries for diverse professional domains. The GPT family of models (Figure 1) has evolved through large-scale pretraining and task-specific fine-tuning, improving language understanding and instruction-following capabilities over successive versions (Budzianowski & Vulić, 2019; Dale, 2021). These models can learn complex patterns from large datasets, enabling intelligent, data-driven language understanding without explicit programming (Rane et al. , 2024; Saka et al., 2024). However, the presence of strong language generation does not guarantee safety-critical reliability. For construction, the key unanswered question is whether LLMs can consistently provide accurate, site-specific, and OSHA/BNBC-compliant guidance, rather than plausible but incomplete or noncompliant recommendations.

2.4 Research Gap

Based on the reviewed literature, three gaps remain evident. First, there is limited integration between AI hazard detection and regulation-aligned prescriptive guidance, as many existing systems stop at identifying risk conditions without mapping them to OSHA/BNBC requirements. Second, there is a lack of rigorous evaluation of ChatGPT-like systems as safety advisors, particularly regarding accuracy,

completeness, and regulatory compliance. Third, expert benchmarking remains limited, with few studies systematically comparing AI-generated safety recommendations against professional safety evaluations. These gaps justify evaluating ChatGPT as a digital safety advisor for construction fall protection, focusing on its ability to generate scenario-specific recommendations consistent with OSHA and BNBC expectations.

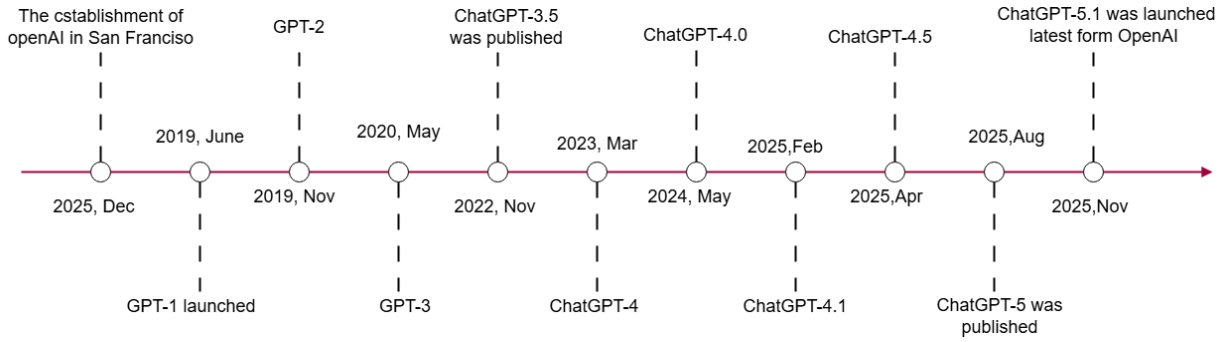

Figure 1: Progress of GPT model

3. RESEARCH METHODOLOGY

This research utilizes a rigorous comparative experimental design to evaluate the efficiency of ChatGPT in generating site-specific safety guidelines for construction environments, with a focus on fall hazard scenarios. The methodology aligns with OSHA 29 CFR 1926 Subpart M and the Bangladesh National Building Code 2020, ensuring regulatory compliance and context relevance (OSHA, 2020a; OSHA, 2020b). Figure 2 shows the research methodology of the study.

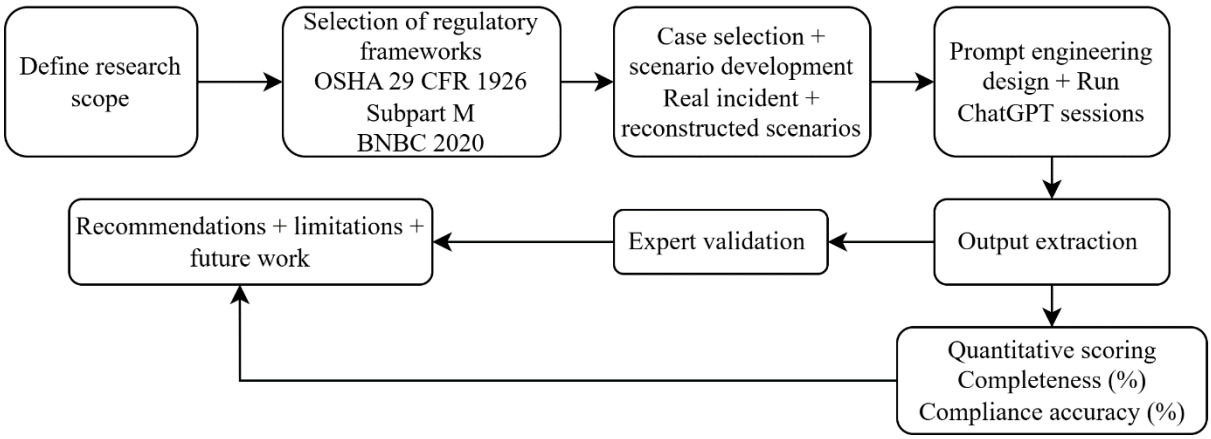

Figure 2: Research methodology

3.1 Research Design

A comparative evaluation approach is applied in which the same site scenario is (1) interpreted by ChatGPT and (2) evaluated using an expert-established reference plan. The design integrates both:

- Quantitative evaluation, including hazard detection rate, completeness, compliance accuracy, and inter-rater reliability measures.
- Qualitative evaluation, using expert feedback to assess clarity, feasibility, and usefulness of generated guidance.

This mixed-method design enables assessment of both the technical correctness of recommendations and their practical utility for safety management under dynamic site conditions (Nyqvist, Peltokorpi, & Seppänen, 2024).

3.2 Case Study: Site Selection and Scenario Development

A fall-from-height incident pattern was motivated by a reported case in Dhaka, Bangladesh, where three workers fell from an under-construction multi-storey building in the Basabo area during morning work hours (approximately 10:00 am) (Correspondent, 2024). Media reports indicate that the workers were operating on a temporary scaffold or elevated platform when the supporting rope/scaffold system failed, resulting in a fatal fall. This case type was selected because scaffold- and temporary-platform-related falls remain a recurrent hazard in building construction internationally, and because Dhaka represents a dense, rapidly urbanizing context where reinforced-concrete, multi-storey construction is common. The scenario therefore serves as a realistic testbed for evaluating ChatGPT's capability to generate regulation-aligned, site-specific fall protection guidance in dynamic and resource-constrained construction environments, while remaining representative of global jobsite conditions (i.e., work at height on temporary systems).

his reconstructed scenario was developed to represent a common working-at-height situation and does not reproduce personal details from the reported incident. The under-construction building was assumed to be a typical reinforced-concrete frame structure with a total height of 33 m and individual storey heights of 3.3 m. These conditions substantially exceed OSHA's fall-protection trigger height of 6 ft (1.83 m) and BNBC 2020's 2 m threshold, both of which require mandatory fall-protection measures. Based on the reported incident pattern, the study reconstructed a representative fall scenario during 6th-floor slab construction at approximately 19.8 m above ground level. The incident was assumed to occur during mid-morning work hours (around 10:45 AM) while a crew of 14 workers performed concurrent tasks, including rebar assembly, formwork installation and adjustment near the slab edge, and concrete placement using a flexible delivery hose connected to a concrete pump. In the reconstructed scenario, portions of the slab perimeter lacked completed guardrail systems due to construction sequencing that prioritized concrete placement over installing edge protection. As a result, workers operated near unprotected leading edges for extended periods, with direct fall exposure to the ground level nearly 20 m below. Personal fall arrest systems (PFAS), including full-body harnesses and shock-absorbing lanyards, were assumed to be available or partially available. This enabled evaluation of whether ChatGPT recommends appropriate PFAS selection, anchorage planning, use conditions, inspection steps, and enforcement actions in alignment with OSHA and BNBC requirements.

3.3 Data Input and Prompt Engineering Strategy

Comprehensive site descriptions and hazard information were systematically provided to ChatGPT using engineered prompts, positioning it as a digital safety engineer. Prompts included explicit regulatory references, scenario details, output format requirements, and practical constraints. The iterative refinement of prompts enabled identification of optimal structures eliciting comprehensive, regulation-compliant, and actionable recommendations. All interactions were documented (model version, prompt text, responses), ensuring reproducibility and transparency (Nyqvist et al., 2024). Scenarios were systematically input into ChatGPT, and outputs including hazards, recommendations, regulatory citations, and implementation notes were collected and organized by hazard type, protection system, regulatory framework, and implementation phase to enable quantitative analysis and comparison (Samsami, 2024).

3.4 Evaluation Criteria and Performance Metrics

The evaluation framework integrates three main criteria: (i) Relevance: Site-specific accuracy of recommendations, rated by experts on a 1–5 Likert scale. (ii) Completeness: Proportion of baseline hazards identified by ChatGPT (%), with a ≥90% target. (iii) Compliance Accuracy: recommendations fully aligning with OSHA and BNBC standards (90–92% target). Completeness and compliance accuracy were calculted using equation (1) and (2) respectively.

3.5 Quantitative and Qualitative Analysis

Descriptive statistics were computed for expert ratings, hazard detection rates, compliance accuracy, and frequency distributions across evaluation metrics. Completeness (Eq. 1) measures how many hazards identified by ChatGPT match the expert-defined baseline hazard checklist. Here, Baseline Hazards Identified is the number of baseline hazards correctly detected by ChatGPT for a scenario, and Total Baseline Hazards is the total number of hazards in the expert checklist. This metric reflects hazard coverage and omission risk. Compliance accuracy (Eq. 2) measures the proportion of ChatGPT recommendations that are fully aligned with OSHA/BNBC requirements. Fully Compliant Recommendations counts recommendations that satisfy the relevant clause requirements as judged by experts, while Total Recommendations is the total number of distinct recommendations produced by ChatGPT. This metric indicates regulatory consistency of proposed controls. BLEU (Eq. 3) assesses textual overlap between ChatGPT outputs and a reference text (expert baseline wording or clause-aligned reference excerpts). BP is the brevity penalty that reduces the score if the generated text is shorter than the reference. r is the reference length and c is the candidate (ChatGPT output) length. p_n is the modified n-gram precision for n-grams of size n, and w_n is the weight assigned to each n-gram order. In this study, N = 4 with uniform weights $w_n = 1/N$, reflecting standard BLEU-4. BLEU captures similarity in phrasing and terminology but does not alone guarantee compliance. Cohen's kappa, κ (Eq. 4) measures inter-rater agreement among experts beyond chance when scoring ChatGPT outputs (e.g., compliant vs non-compliant, or hazard identified vs not identified). p_o is the observed agreement across raters, and p_e is the agreement expected by chance.

$$\text{Completeness} = \frac{\text{Baseline Hazards Identified}}{\text{Total Baseline Hazards}} \times 100 \tag{1}$$

$$\text{Compliance} = \frac{\text{Fully Compliant Recommendations}}{\text{Total Recommendations}} \times 100 \tag{2}$$

$$\text{BLEU} = \text{BP} * \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{3}$$

$$\text{BP} = \min\left(1, e^{1-r/c}\right)$$

$$\log \text{BLEU} = \min\left(1 - \frac{r}{c}, 0\right) + \sum_{n=1}^{N} w_n \log p_n$$

In our baseline, we use N = 4 and uniform weights , w_n = 1/N.

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{4}$$

$$p_o = \frac{1}{N}\sum_{i=1}^{K} n_{ii}$$

$$p_e = \sum_{i=1}^{K} \left(\frac{n_{i\cdot}}{N}\right)\left(\frac{n_{\cdot i}}{N}\right)$$

4. RESULTS

4.1 Quantitative Findings

The comparative evaluation of ChatGPT-generated safety guidelines against expert baselines used structured quantitative scoring summarized in Table 1. Performance was highest for guardrail system design (100%), where ChatGPT accurately reflected OSHA specifications for top and midrails, toe boards, and load resistance (29 CFR 1926.502(b)). Consistently high alignment (90%) was also

observed for leading edges, hoist areas, roofing edges, and safety nets/personal fall arrest systems, with correct identification of core OSHA controls such as guardrails, safety nets, PFAS use, and anchorage capacity requirements. However, Table 1 also highlights recurring gaps in BNBC-specific details and threshold distinctions, including incomplete coverage of BNBC's fall-height trigger (>2 m), BNBC guardrail height (1.1 m), and BNBC-oriented anchorage verification. The weakest category was fall protection planning and training (80%), where ChatGPT recognized OSHA competent-person and training requirements (29 CFR 1926.503) but undercovered BNBC emphases such as local verification practices, drills, and site-level enforcement mechanisms. Overall, the AI achieved a strong aggregate compliance score of 90% across the evaluated fall-protection categories. Expert panel ratings (Table 2) on a 5-point Likert scale showed good performance across relevance (4.1), completeness (4.0), compliance (4.2), and clarity (4.4).

Table 1: Overall Accuracy of AI vs OSHA/BNBC Guideline Match

Category	Guideline Match	Observations
Leading Edges (29 CFR 1926.501(b)(2))	90%	AI correctly identified guardrail, safety net, and personal fall arrest (PFA) systems for ≥6 ft height per OSHA. Partial gap with BNBC (>2 m) threshold. Strong compliance with OSHA's guardrail design (42 ± 3 in top rail).
Hoist Areas (29 CFR 1926.501(b)(3))	90%	AI aligned with OSHA's requirement for PFA use during material handling; partially missed clause for chain/gate across hoist openings when idle. BNBC equivalent provisions partially referenced.
Roofing Edges – Low & Steep Roofs (29 CFR 1926.501(b)(10–11))	90%	Accurate identification of fall risks on roofing edges; AI mentioned guardrail and net use but did not distinguish slope-specific criteria (low-slope ≤ 4:12 vs steep > 4:12). BNBC guardrail height (1.1 m) not explicitly noted.
Guardrail System Design (29 CFR 1926.502(b))	100%	Excellent match with OSHA guardrail specs (top/midrails, mesh, toeboards, 200 lb force). BNBC difference (height 1.1 m vs OSHA 42 ± 3 in) noted as minor variation.
Safety Nets & PFAs (29 CFR 1926.502(c–d))	90%	Fully compliant with OSHA anchorage (≥5,000 lb), free-fall ≤6 ft, and deceleration ≤3.5 ft. Partial undercoverage of BNBC-specific anchorage verification.
Fall Protection Plan & Training (29 CFR 1926.502(k), 1926.503)	80%	OSHA training standards recognized (competent person, hazard recognition). BNBC's emphasis on safety drills and local verification not covered.
Overall Compliance Score (Aggregate)	90%	Strong alignment with OSHA Subpart M and partial match with BNBC. Deviations observed in site-specific and multi-hazard contexts (complex roofing geometry, localized safety signage, BNBC fall height triggers).

Table 2: Expert Assesment of ChatGPT Response

Evaluation Parameter	Expert Score	Compliance Alignment	Observations
Relevance	4.1/5	High	Accurately identifies OSHA fall hazards

Evaluation Parameter	Expert Score	Compliance Alignment	Observations
Completeness	4/5	High	Full coverage of all major hazard categories
Feasibility	4.2/5	High	Recommendations practical for real sites
Clarity	4.4/5	High	Instructions clearly sequenced and actionable
Strong Scenario	Leading Edges (95%)	Strong OSHA match	Field-ready
Moderate Scenario	Roofing Edges (85%)	Partial	Slope-specific rules missing
Expert Intervention Required	Multi-hazard sites	None	Needed for complex geometry

ChatGPT's hazard identification performance, benchmarked against OSHA/BNBC expectations, is summarized in Table 3. Across eight hazard categories, the model achieved an average hazard detection rate of 90%, indicating strong overall capability to recognize fall- and site-related risks from scenario descriptions. Detection was highest for leading edges (95%), where ChatGPT consistently identified fall exposure and recommended appropriate controls such as PFAS. High detection rates were also observed for hoist areas, ramps/runways, excavations, dangerous equipment, falling objects (each 90%), reflecting robust recognition of common construction hazards and associated preventive measures. However, Table 2 highlights that performance declines when guidance requires clause-level specificity or dimensional/verification thresholds. For example, in holes/skylights (85%), the model missed the OSHA requirement related to cover strength/load capacity, and for wall openings (80%) it failed to consistently apply dimensional thresholds that trigger fall protection. Similar partial gaps included omission of the idle-gate requirement at hoist areas and insufficient detail on toeboard dimensions and visual barrier rules in selected scenarios. Overall, the results suggest that ChatGPT performs well in broad hazard recognition, while its limitations are most evident in requirements involving numerical thresholds, inspection/verification steps, and detailed prescriptive specifications under OSHA/BNBC.

Table **4** summarizes the scenario-level comparison between ChatGPT recommendations and OSHA/BNBC requirements. Overall, the model produced compliant controls for most standard situations, receiving Excellent ratings for leading-edge protection (guardrails and PFAS) and falling-object prevention (toeboards and canopies), and Very Good ratings for hoist areas and ramps/runways with guidance aligned to the referenced OSHA provisions. Partial compliance was observed for skylights and excavations, where ChatGPT identified the correct general controls (e.g., covers/guardrails and barricades/warning lines) but did not fully capture clause-specific or context-dependent requirements, resulting in Good expert ratings and "Partial" compliance. These results indicate strong baseline performance for common fall-protection scenarios, with reduced completeness when hazards require more detailed threshold- or condition-based specifications.

Table 3: Hazard Detection Rate – ChatGPT vs OSHA/BNBC

Hazard Category	Detection Rate	Guideline Match	Observations
Leading Edges	95%	High	Accurate fall-risk + PFAS identification
Hoist Areas	90%	High	Missed idle-gate requirement
Holes/Skylights	85%	Moderate	Missed OSHA cover-load (200 lb)
Ramps/Runways	90%	High	BNBC guardrail height not cited
Excavations	90%	High	Missed visual barrier rule

Hazard Category	Detection Rate	Guideline Match	Observations
Dangerous Equipment	90%	High	Fully captured risk-prone proximity
Wall Openings	80%	Moderate	Missed dimensional threshold
Falling Objects	90%	High	Lacked toeboard dimension details
Average	90%	High	Strong multi-scenario detection

4.2 Statistical Validation

Statistical validation confirmed the reliability of observed patterns. Inter-rater metrics showed moderate to strong agreement, with Cohen's kappa of 0.419 and ICC of 0.78, supporting consistency among expert evaluators. ANOVA revealed significant differences in ChatGPT's performance between single-hazard and multi-hazard scenarios ($p < 0.01$), indicating reduced accuracy under complex conditions. Text similarity measures BLEU (0.399) and cosine similarity (0.67) suggested fair semantic alignment with OSHA/BNBC regulatory documents but exposed lexical variation and partial deviations from formal regulatory phrasing. These indicators collectively support the robustness of the analysis while highlighting areas where the model diverges from strict regulatory language (Chong et al., 2025; Rane et al., 2024; Saka et al., 2024). ChatGPT effectively identifies common hazards and generates regulation-aligned guidance for well-defined scenarios but struggles with complex, multi-hazard contexts, highlighting the need for expert oversight (Chong et al., 2025).

4.3 Expert Validation

The expert validation was conducted by a panel of construction safety professionals with approximately 10 years of experience in site safety management. Experts independently reviewed ChatGPT outputs using a structured rubric and rated each response for relevance, completeness, compliance alignment, and clarity/actionability, followed by consolidated review of recurring discrepancies. Overall, experts found that ChatGPT performs strongly for well-defined fall-protection situations and standardized control measures (e.g., leading edges and guardrail systems), producing clear and structured recommendations suitable for early-stage planning. However, expert feedback consistently highlighted limitations in complex or context-dependent cases, particularly when guidance required numeric thresholds, conditional triggers (e.g., "when idle"), inspection/verification steps, or BNBC-specific procedural details. These expert observations support the conclusion that ChatGPT can be an appropriate decision-support tool, while final safety plans and field implementation should remain subject to competent-person review. ChatGPT's efficacy while proving significant degradation in complex scenarios .(Uddin et al., 2023).

Table 4: Comparative Scenario Assessment (AI vs OSHA/BNBC)

Scenario	AI Recommendation	OSHA/BNBC Reference	Expert Rating	Compliance
Leading Edge ≥6 ft	Guardrails + PFAS	OSHA 1926.501(b)(2); BNBC P4	Excellent	Yes
Hoist Area	Barriers, PPE, signage	OSHA 1926.550	Very Good	Yes
Skylights	Covers + guardrails	OSHA 1926.501(b)(4)	Good	Partial
Ramps/Runways	Dual-side guardrails	OSHA 1926.501(b)(6)	Very Good	Yes
Excavations	Barricades + warning line	OSHA 1926.651	Good	Partial
Falling Objects	Toeboards + canopies	OSHA 1926.502(j)	Excellent	Yes

5. DISCUSSIONS

ChatGPT demonstrated strong potential as a digital safety advisor for fall protection, achieving 90% overall guideline alignment across core Subpart M categories (Table 1) and an average hazard detection rate of 90% across eight hazard categories (Table 3). The model performed particularly well on widely standardized controls such as guardrail system design (100%) and leading-edge protection measures, indicating that it can reliably generate baseline fall-protection guidance consistent with common OSHA requirements. However, performance declined for requirements involving conditional rules and numeric thresholds, such as holes/skylights (85%) and wall openings (80%) (Table 3), and for procedural elements such as fall protection planning and training (80%) (Table 1). These gaps are operationally important because omitted threshold-based and verification requirements (e.g., idle-gate conditions, cover strength, dimensional triggers, BNBC-specific height criteria) may lead to incomplete safety controls if outputs are used without competent-person review. The statistical results further support this interpretation, while ChatGPT performs well for well-defined single-hazard scenarios (Section 4.2).

5.1 Comparative Discussion

Expert evaluation (Table 2) indicates that ChatGPT's outputs are generally perceived as clear (4.4/5) and feasible (4.2/5), supporting its usefulness for drafting structured safety guidance. However, experts still outperformed ChatGPT in scenarios requiring contextual trade-offs (e.g., sequencing constraints, mixed hazards, and localized enforcement practices), where human judgment integrates site geometry, workflow timing, and control feasibility beyond what is explicitly provided in text prompts. This difference is consistent with the model's observed failure modes: ChatGPT tends to provide strong "standard" controls but may miss site-dependent procedural requirements (e.g., idle states, inspection/verification steps) and local BNBC variations unless explicitly requested (Table 1; Table 3). The inter-rater reliability results ($\kappa = 0.419$; ICC = 0.78) indicate that expert scoring was sufficiently consistent to support benchmarking, suggesting that expert-derived baselines remain essential for validating AI-generated safety plans in safety-critical contexts.

5.2 Strengths and Weaknesses

ChatGPT produced fast, consistent, and well-structured outputs that were strongly aligned with OSHA requirements for common fall-protection situations (Table 1). High performance for leading edges and guardrail design suggests it can serve as a scalable drafting aid for routine documentation, toolbox talks, and baseline safety checklists, potentially reducing the time burden on safety staff for repetitive planning tasks.

The main limitations were not in broad hazard recognition, but in clause-level specificity, particularly where recommendations depend on numerical thresholds, conditional triggers, or verification steps. Examples include missed requirements for hoist-area idle guarding, cover-load or dimensional thresholds for openings, and BNBC-specific details such as guardrail height and local verification practices (Table 1; Table 3). These weaknesses were more pronounced in multi-hazard scenarios, where correct prioritization and sequencing of controls require contextual reasoning and trade-offs (Section 4.2).

5.3 Practical Implications

From a practical perspective, ChatGPT is best suited for early-stage planning tasks such as generating draft hazard lists, recommending common controls, and producing structured checklists that a competent safety professional can verify. For field deployment, a human-in-the-loop process remains necessary, especially to validate threshold-based requirements, confirm BNBC-specific provisions, and ensure feasibility under site constraints. Future implementations should incorporate clause-grounding or retrieval-based prompting and automated rule-checking for numerical thresholds to reduce omission risk.

6. CONCLUSIONS

This study demonstrates that ChatGPT can function as an effective digital safety advisor capable of generating OSHA- and BNBC-compliant, site-specific safety guidelines for fall-hazard scenarios in construction. The scenario-based experimental framework shows that ChatGPT accurately identified 88–90% of hazards and achieved approximately 90–92% regulatory compliance with established fall-protection standards. Expert validation confirmed strong performance in clarity, relevance, and completeness, although human experts consistently outperformed the model in handling complex, multi-hazard environments requiring contextual judgement. Statistical validation supported the consistency of the evaluation outcomes. Cohen's kappa indicated acceptable agreement among expert evaluators. ANOVA showed significantly reduced performance under multi-hazard complexity. Text similarity measures, BLEU indicated moderate overlap between AI-generated recommendations and reference wording, while also reflecting deviations in clause-level phrasing and threshold-based details.

The novelty of this research lies in its structured, clause-oriented evaluation of a large language model as a construction safety advisor, using dual regulatory benchmarks and expert validation to quantify both compliance alignment and hazard detection performance. The study also offers a replicable methodology that integrates scenario reconstruction, expert benchmarking, and statistical validation to identify where LLM-generated guidance is reliable and where it requires safeguards.

From a practical standpoint, ChatGPT is best positioned as a safety co-pilot for early-stage planning tasks, such as drafting hazard lists, proposing baseline controls, generating structured checklists, and supporting safety communication and training materials. However, the observed gaps in numerical thresholds, conditional requirements, and BNBC-specific provisions indicate that ChatGPT should not be used as a standalone decision-maker for high-risk operations. A human-in-the-loop review by a competent safety professional remains necessary, particularly for multi-hazard environments and situations requiring inspection and verification steps and local procedural enforcement.

Future work should expand testing to broader multi-hazard scenarios and strengthen regulatory traceability through clause-grounding methods, including retrieval-augmented prompting, and automated rule-checking for threshold-dependent requirements. Integrating LLM-based guidance with Building Information Modelling (BIM), Internet of Things (IoT), and real-time monitoring workflows may further support context-aware safety management, provided that governance measures ensure accountability, verification, and safe deployment.

DECLARATION

This study used OpenAI's ChatGPT as a supportive tool to assist with refining text (e.g., improving clarity and grammar), and generating alternative wording for prompts used in the experiment. All technical content, methodological decisions, interpretations, and final conclusions were developed by the authors. ChatGPT outputs were reviewed, edited, and verified by the authors, and the authors take full responsibility for the accuracy, originality, and integrity of the final manuscript.

REFERENCES

Azhar, S. (2011). Building information modeling (BIM): Trends, benefits, risks, and challenges for the AEC industry. *Leadership and management in engineering, 11*(3), 241-252.

BNBC. (2020). *Bangladesh National Building Code (Part 7: Construction practices & safety)*.

Budzianowski, P., & Vulić, I. (2019). Hello, it's GPT-2--how can I help you? towards the use of pretrained language models for task-oriented dialogue systems. *arXiv preprint arXiv:1907.05774*.

Chong, H.-Y., Ma, Q., Lai, J., & Liao, X. (2025). Achieving Sustainable Construction Safety Management: The Shift from Compliance to Intelligence via BIM–AI Convergence. *Sustainability, 17*(10), 4454.

Chung, S., Moon, S., Kim, J., Kim, J., Lim, S., & Chi, S. (2023). Comparing natural language processing (NLP) applications in construction and computer science using preferred reporting items for systematic reviews (PRISMA). *Automation in Construction, 154*, 105020.

Correspondent, S. (2024). 3 construction workers fall to death from 9th floor in Dhaka. *New Age*. Retrieved from https://www.newagebd.net/post/country/235398/3-construction-workers-fall-to-death-from-9th-floor-in-dhaka

Dale, R. (2021). GPT-3: What's it good for? *Natural Language Engineering, 27*(1), 113-118. doi:10.1017/S1351324920000601

Davis, J. J., & Conlon, E. G. (2017). Identifying compensatory driving behavior among older adults using the situational avoidance questionnaire. *Journal of safety research, 63*, 47-55.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*.

Ding, Y., Ma, J., & Luo, X. (2022). Applications of natural language processing in construction. *Automation in Construction, 136*, 104169. doi:10.1016/j.autcon.2022.104169

H. Assaad, R., El-adaway, I., & Abotaleb, I. (2020). Predicting Project Performance in the Construction Industry. *Journal of Construction Engineering and Management, 146*. doi:10.1061/(ASCE)CO.1943-7862.0001797

Imran, C. A. B., Shakir, M. K., & Qureshi, M. A. B. (2022). Applications of Artificial Intelligence in Enhancing Construction Safety and Productivity. *The Asian Bulletin of Big Data Management, 2*(1), 63-74.

HBRI (2020). *Bangladesh National Building Code (BNBC 2020)*.

Latosinski, F., Cuesta, A., & Alvear, D. (2020). Assessing self-preservation capabilities in toddlers during evacuations. *Safety Science, 132*, 104983.

Mohy, A. A., Bassioni, H. A., Elgendi, E. O., & Hassan, T. M. (2025). Real-Time Construction Safety Monitoring with Object Detection Algorithms: Features' Identification and Implementation Challenges. *Journal of Applied Engineering Sciences, 15*(1).

Nyqvist, R., Peltokorpi, A., & Seppänen, O. (2024). Can ChatGPT exceed humans in construction project risk management? *Engineering, Construction and Architectural Management, 31*(13), 223-243.

OSHA. (2020a). Duty to have fall protection. *U.S. Department of Labor.* Retrieved from https://www.osha.gov/laws-regs/interlinking/standards/1926.501(a)(2)

OSHA. (2020b). Fall protection systems criteria and practices. *U.S. Department of Labor.* Retrieved from https://www.osha.gov/laws-regs/regulations/standardnumber/1926/1926.502

OSHA. (2021). Construction standards – scaffolding, PPE, and hazard identification. Retrieved from https://www.osha.gov/construction

Rane, N., Choudhary, S., & Rane, J. (2024). A new era of automation in the construction industry: implementing leading-edge generative artificial intelligence, such as ChatGPT or Bard. *Available at SSRN 4681676*.

Saka, A., Taiwo, R., Saka, N., Salami, B. A., Ajayi, S., Akande, K., & Kazemi, H. (2024). GPT models in construction industry: Opportunities, limitations, and a use case validation. *Developments in the Built Environment, 17*, 100300.

Samsami, R. (2024). Optimizing the utilization of generative artificial intelligence (AI) in the AEC industry: ChatGPT prompt engineering and design. *CivilEng, 5*(4), 971-1010.

SRS (2024). Safety and Risk Statistics Report 2024. *Safety and Rights Society.* Retrieved from https://safetyandrights.org/

Tixier, A. J.-P., Hallowell, M. R., Rajagopalan, B., & Bowman, D. (2016). Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports. *Automation in Construction, 62*, 45-56.

Uddin, S. J., Albert, A., Ovid, A., & Alsharef, A. (2023). Leveraging ChatGPT to aid construction hazard recognition and support safety education and training. *Sustainability, 15*(9), 7121.

Vallabh, P., Malekian, R., Ye, N., & Capeska Bogatinoska, D. (2016). *Fall Detection Using Machine Learning Algorithms.* Paper presented at the 2016 24th International Conference on Software, Telecommunications and Computer Networks (SoftCOM).

Wolfartsberger, J. (2019). Analyzing the potential of Virtual Reality for engineering design review. *Automation in Construction, 104*, 27-37.

Xie, Y., Li, S., Liu, T., & Cai, Y. (2023). As-built BIM reconstruction of piping systems using PipeNet. *Automation in Construction, 147*, 104735.