

Comparative Analysis of Cognitive Agreement between Human Analysts and Generative AI in Construction Safety Risk Assessment

Unmesa Ray¹, Cristian Arteaga² and JeeWoong Park³

^{1,2,3} Department of Civil and Environmental Engineering and Construction, University of Nevada, Las Vegas, USA

rayu1@unlv.nevada.edu, arteagas@unlv.nevada.edu, jee.park@unlv.edu

Abstract -

The construction industry struggles with safety risk assessment complexities due to evolving work environments, diverse labor forces, time constraints, regulatory intricacies, and inconsistent practices. While previous studies have highlighted the potential of Artificial Intelligence (AI) in automating processes and enhancing safety assessment, a gap exists in the convergence between human analysts and language AI models. Therefore, this study assesses the alignment in identifying risk factors by human analysts and a Language Model (LM) in Occupational Safety and Health Administration (OSHA) accident reports. Furthermore, it: 1) categorizes error types, 2) establishes an acceptance threshold for LM-generated responses, and 3) evaluates inter-rater reliability in construction accident content analysis. The test results reveal significant convergence between human and machine responses and identify potential hallucination effects in generative AI, thus paving the way for improved safety risk assessments within the construction industry.

Keywords –

Construction industry, safety risk assessment, Artificial Intelligence (AI), Occupational Safety and Health Administration (OSHA), Language Model (LM), inter-rater reliability, generative AI

1 Introduction

1.1 Motivation

The construction industry presents a significant problem concerning workplace accidents, primarily due to its inherently high-risk nature. Despite considerable

efforts over the past several years, the safety aspects of the construction industry have not witnessed enhancements at a level comparable to that observed in other industries. As a result, the industry still suffers with a substantial number of accidents [1]. In 2021, the construction industry accounted for a staggering 21% of all occupational fatalities in the United States, as reported by the Bureau of Labor Statistics [2]. 65.5% of these construction-related deaths are attributed to the "Focus Four" incidents, including falls, which accounted for 35% of the fatalities, struck-by incidents at 17%, electrocutions at 7.6%, and caught incidents at 5.8% [3].

Construction safety issues have seen a significant decline in injuries and fatalities after the implementation of OSHA's rules and standards [4]. All employers are required to notify OSHA upon death, injury, or hospitalization. The accident report form is available on OSHA's website. Employers may also contact a designated phone number to report accidents. The form contains structured data such as accident date, company name, and age/sex of injured person, as well as a narrative describing the accident. This accident narrative provides valuable information to better understand the context and sequence of events that led to the incident.

However, investigation summary, due to its unstructured nature, and possible variations of descriptive context poses several challenges in analysis and interpretation. While quantitative data is generally straightforward to handle, the unstructured and descriptive nature of narratives in accident reports presents challenges for systematic analysis. Figure 1 shows a section of OSHA Form 301 that requires investigation summary.

14)* **What was the employee doing just before the incident occurred?** Describe the activity, as well as the tools, equipment, or material the employee was using. Be specific. Examples: "climbing a ladder while carrying roofing materials"; "spraying chlorine from hand sprayer"; "daily computer key-entry."

15)* **What Happened? Tell us how the injury occurred.** Examples: "When ladder slipped on wet floor, worker fell 20 feet"; "Worker was sprayed with chlorine when gasket broke during replacement"; "Worker developed soreness in wrist over time."

Figure 1. Segment for investigation summary details in OSHA's Form 301

Consequently, the manual analysis of these narratives is a time-consuming process, which ultimately translates into substantial labor. A previous study [5] introduced framework that facilitated the extraction of standard risk factors and outcome variables from OSHA accident reports, providing a comprehensive approach to addressing the limitations of accident data analysis. However, the analysis of such databases demands highly skilled labor, making it an expensive and time-consuming administrative method [6]. Likewise, traditional content analysis has historically relied on human analysts, incorporating inter-rater reliability measures into the coding system. On the other hand, a class of neural networks known as Transformers [7] has significantly streamlined language classification tasks in recent years. However, before fully embracing their potential, it is important to assess their alignment with human coders, particularly in the realm of content analysis.

1.2 Related Work

Researchers have adopted numerous approaches to effectively handle and covert database into useful information [8]. However, one notable drawback of these methods is their lack of reliance on empirical data, coupled with a limited scope of application [9], [10]. In response to these limitations, a unified attribute-based framework has been proposed [need a citation]. This framework enables the extraction of standard risk factors and outcome variables from naturally occurring accident reports, providing a comprehensive approach to

addressing the shortcomings in accident data analysis [5]. Due to the high dimension of the injury report feature space and the diversity of construction situations, the available training data are naturally sparse. Therefore, manually analyzing tens of thousands of injury reports would have been required to put together a satisfactory training database and achieve efficient machine learning [11].

Previously, a study [12] has proposed an analytical method incorporating text mining and interpretable machine learning to discern factors influencing injury severity levels within traffic crash narratives. However, it suffered from a limitation in the text analysis since it was based on word frequency rather than meaningful semantics from the narrative. On the contrary, LMs have gained significant attention in recent years due to their ability to generate human-like text and perform a wide range of language-based tasks. In the construction industry, LMs have the potential to improve efficiency, accuracy, and communication in several different ways [13]. Regarding language representation models, different approaches have been developed in the past decade [14]. One major model is the autoregressive language model Generative Pre-trained Transformer (GPT). Developed by OpenAI, GPT is trained on a large dataset of text and can generate human-like text. GPT 3.5 has been fine-tuned for information retention during the conversation, making it suitable for activities that are solely reserved for human interaction [15]. A recent study used GPT to generate a construction schedule for a simple construction project and the outcome shows that it can generate a coherent schedule that follows a logical approach to fulfill requirements of the scope indicated [13]. Another study integrated GPT as an intervention to aid hazard recognition efforts in the curriculum of students pursuing a career in the construction industry. The results suggest that GPT can be leveraged to improve hazard recognition levels [15]. To the best of our knowledge, this study marks the first attempt to assess the alignment of Transformer models with human analysts in comprehensive understanding of accident data, offering a promising avenue for future research and applications in the field of accident analysis and prevention.

1.3 Research Objective and Scope

This study's primary objective is to assess the alignment in identifying contributing risk factors in accidents as perceived by human analysts versus an LM. By undertaking this comparative analysis, the research objectives are threefold: (1) defining a threshold for accepting LM-generated responses (2) systematically categorizing error types into Type I and Type II, and (3) evaluating the inter-rater reliability between human

analysts and the LM in the context of content analysis within construction accidents. This study implements the proposed approach by analyzing 150 accident investigation summaries related to "Fall from Heights." These summaries are sourced from the official website of OSHA.

2 Methodology

Figure 2 is an illustration of the steps adopted in this study, which include a data collection process, manual analysis of narratives, the setup of LM for generating machine responses, and the integration of a suitable statistic within the comparison of LM's responses against the manually identified contributing factors.

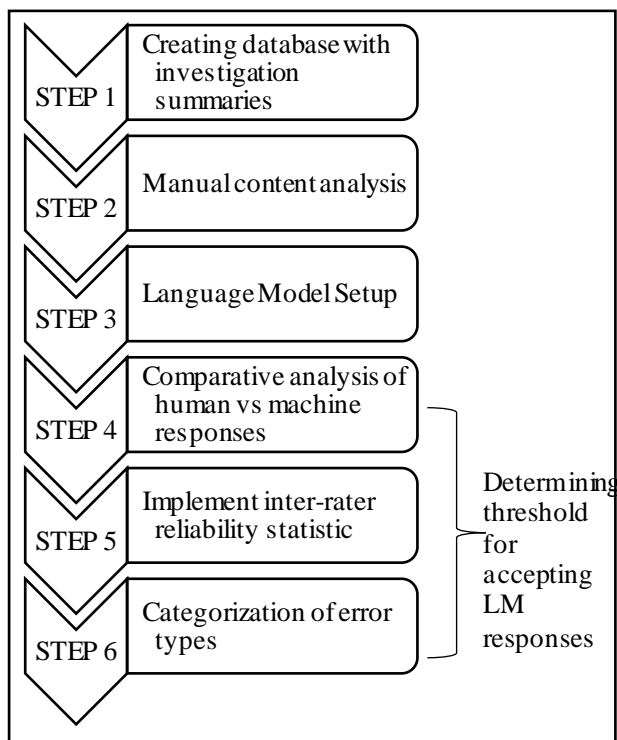


Figure 2. Proposed research methodology

2.1 Creating Database with Investigation Summaries

Initiating the proposed framework involved retrieving accident report narratives, which constitute unstructured data available on OSHA's official website, followed by the utilization of simple and easily accessible Microsoft Office tools, such as Excel, to construct a database cataloging accident identification code (IDs) alongside their respective narratives. This

ensures establishing an organized repository for subsequent content analysis.

2.2 Manual Content Analysis

Content analysis, critical to this research, involves identifying risk factors contributing to accidents within narratives. In the context of this study, a risk factor can be defined as a specific element, condition, or circumstance identified in accident reports that contributes to the occurrence or severity of accidents. These factors encompass various aspects such as environmental conditions, human behaviors, equipment usage, or any other variable that plays a role in the initiation or escalation of accidents. Recognizing these factors necessitates pattern recognition through manual labor—interpreting and inferring accident causes. We enlisted and trained a dedicated researcher for this analysis. The identified risk factors by the human analyst establish the ground truth for determining the acceptance threshold for LM-generated responses.

2.3 LM Setup and Analysis

In this phase, we integrated GPT with the collected data. The API key obtained from OpenAI facilitated interfacing Python, thereby, aiding to input prompts into the LM. The accuracy of information extraction depends on the prompt. Hence, a meticulously devised prompt must include desired outcomes in structured format, while excluding specific redundant details. We also consider fine-tuning the prompt for enhanced information retrieval.

2.4 Comparative Analysis of Responses – Human versus Machine

Upon receiving responses from the LM, we conducted a comparative analysis between contributing risk factors identified by the human analyst and the machine. A manual review of all accident reports was performed, comparing factors for each narrative. The researcher evaluating the comparative analysis considers the LM responses valid only when the LM's response aligns with the true cause leading to the accident. We treated the LM as an independent coder akin to the human analyst, and percent agreement was used to measure inter-coder reliability. Percent agreement simply represents number of agreements over total number of measures [16]. The formula for percent agreement is given below where A represents the number of agreements between the machine and the human analyst, and N is the total number of factors identified by both.

$$PA_o = A/N \quad (1)$$

2.5 Error Type Formulation and Implementing Inter-rater Reliability Statistic

In the concluding phase, we categorize the types of errors that may occur. When comparing human versus machine responses, three scenarios are possible. The first involves complete agreement, where both human and machine responses are identical, resulting in 100% agreement. However, in other cases, Table 1 shows the errors that may occur and can be classified as follows:

- **Type I Errors:** Occur when human responses include certain risk factors that are absent in the LM-generated responses.
- **Type II Errors:** Occur when the LM-generated responses include certain factors that the human analyst does not include.

Table 1. Formulation of Type I and Type II errors

		Manual Coding	
Machine Coding	Identified	Pass	Type II Error
	Not Identified	Type I Error	Pass

3 Case Study

The proposed methodology leveraged OSHA accident reports as the primary dataset. Four researchers were hired and trained for data collection and manual analysis, working 6 hours a day for four weeks. This entails a thorough manual review process, where the researchers sift through accident reports to identify 150 cases resulting in fatal injuries specifically caused by falls from heights. Additionally, each researcher agrees on the contributing risk factors. In instances where discrepancies arise, the researchers conducted re-reviews to ensure accuracy until a consensus was reached. Discussions are held to establish consistent terminology. This phase is crucial as the manual coding process aims to ensure consistency and establish the ground truth for the subsequent comparative analysis with machine-generated responses. Upon examination, roof work emerges as a prevalent cause. After curating the relevant narratives, a database was constructed and integrated into Python for a comprehensive analysis using the LM. Subsequently, prompts were tailored to meet the specific requirements of this research. The fine-tuning phase is instrumental in determining the final prompt for utilization. The process of formulating the prompt involves a straightforward brute force method to refine the prompt after each response, aligning the output with

our desired outcome. These prompts are generalizable across domains, contingent on the integration of the brute force method into the process. This ensures that users can interact with the LM to extract meaningful information without requiring a background in NLP. The following is the evolution of prompts with the corresponding LM responses, and the encountered challenges throughout this iterative process until the final prompt was decided.

Prompt 1: Please give the risk factors associated with construction safety in the following narrative.

LM response for prompt 1: In the given narrative, various risk attributes associated with construction safety are evident - Working at Height: The employee was on the roof, which is a high-risk location in construction due to the potential for falls and other hazards associated with elevated work.

... seven more attributes with added explanations.

Limitations of prompt 1: Overly detailed factors, redundant OSHA data, and lengthy hazard mitigation paragraphs in tested narratives.

Prompt 2: Please give the risk factors associated with construction safety in the following narrative with no additional sentences and headers.

LM response for prompt 2: Working at Height: The employee was on the roof, which is a high-risk location in construction due to the potential for falls and other hazards associated with elevated work.

... seven more attributes with added explanations.

Limitations of prompt 2: Despite the request to remove descriptive attribute headers, they continue to be present in the responses.

Prompt 3: Please give the risk factors associated with construction safety in the following narrative excluding date, time, company name, age with no additional sentences and remove any headers the attributes begin with.

LM response for prompt 3:

Working at height

Ladder use

Dizziness

Brief rest

Returning to work

Fall from roof

Severe head, neck, and back trauma

Since implementing the final prompt involves the generation of contributing risk factors without an additional description, this prompt is applied to the entire dataset. In the concluding phase, a comprehensive comparative analysis is conducted between LM-generated results and the manually identified contributing factors for each narrative. This qualitative analysis reveals distinct trends as follows:

- Machine responses precisely align with human-identified risk factors.
- The machine identifies all essential accident-contributing factors and introduces additional meaningful attributes, though not necessarily contributing to the accident.
- LM responses fall short in identifying one or more key accident-contributing factors.

4 Results and Discussion

In Figure 3, 104 cases out of the 150 datasets demonstrated an exact match between human-analyzed factors and machine responses. Only 1 case exhibited a Type I error, while 45 cases showed Type II errors. Further, we computed percent agreement for both error types, resulting in 66.67% for Type I errors and an average of 28% for Type II errors.

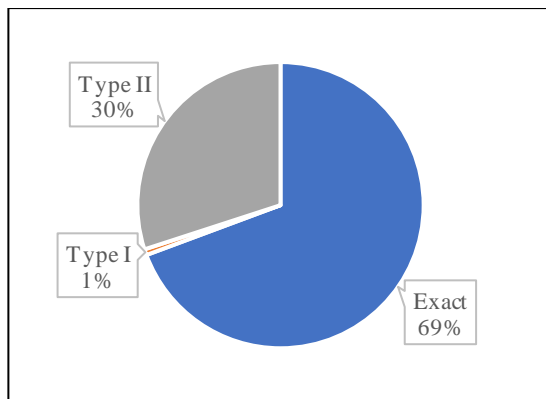


Figure 3. Distribution of alignment between manual versus machine check across the dataset

Reliable coding is achieved when the percent agreement surpasses 70% [17]. Upon examination of the statistics, it becomes evident that human and machine alignment is robust for most cases in our dataset, boasting 100% agreement. However, for error types, the percentage agreement falls below the 70% threshold. Consequently, based on the proposed methodology, it is reasonable to conclude that the inter-rater reliability threshold for human and LM is achieved at 70%.

For Type II errors, a deeper investigation into factors missing in human coding reveals that 86.67% of the

omissions pertain to a broader description of the narrative, which human analysts do not consider as important risk factors contributing to the accident. However, the remaining 13.33% encompasses factors that are not present in the narrative. This suggests a tendency for the LM to introduce additional factors, indicative of a phenomenon referred to as the "hallucination effect" of generative AI.

Example of Type II error that showcases hallucination effects of generative AI:

Investigation Summary: On December 4, 2006, Employee #1 was working for a general contractor that specialized in the construction of single-family houses. He was on a scaffold, when he fell approximately 30 feet to the ground and was killed.

LM Response:

- Employee conducting work at a height on a scaffold.
- Scaffold did not provide sufficient fall protection measures.
- Lack of proper safety equipment or personal protective gear.
- Inadequate training or lack of knowledge on scaffold safety.
- Failure to comply with safety regulations or standards.
- Absence of safety inspection or assessment prior to work.
- Potential negligence in ensuring scaffold stability.
- Insufficient communication or coordination regarding safety precautions.

For the above example, within LM responses, points 2 to 8 are absent in the original narrative.

Example of Type II error that showcases descriptive factors that do not directly contribute to the accident:

Investigation Summary: At approximately 9:00 a.m. on September 13, 2006, an employee was removing concrete forms along a floor opening, on the third floor of a building under construction. While working around the unguarded floor opening, the employee fell approximately 18 ft through the opening and landed on the first floor. The employee was not wearing fall protection. The employee was transported to a local trauma center, where he was pronounced dead on arrival.

LM Response:

- Employee removing concrete forms.
- Unguarded floor opening.
- Employee fell approximately 18 ft.
- Employee landed on the first floor.
- Employee not wearing fall protection.

- Employee transported to a local trauma center.
- Employee pronounced dead on arrival.

For the above example within LM responses, points 3,4,6, and 7 are details of the accident that do not directly contribute to it.

5 Conclusion

The study comprehensively explores the alignment between human analysts and a LM, specifically a Transformer model, in identifying risk factors contributing to construction accidents. For the proposed methodology, we established an acceptance threshold for LM-generated responses, while systematically categorizing error types into Type I and Type II discrepancies for 150 OSHA accident investigation summaries. The comparative analysis between manual and machine responses revealed that 69% of the cases demonstrated an exact match between human-analyzed factors and LM responses with only 1% showing Type I errors and 30% exhibiting Type II errors. Additionally, the calculated percent agreement for Type I and Type II errors were 66.67% and 28%, respectively. While these percentages fall slightly below the 70% threshold suggested for reliable coding, the study underscores the considerable potential of collaborative human and LM analysis, particularly in categorizing error types and establishing thresholds. Furthermore, the hallucination effect underscores the need for a nuanced interpretation of the results generated by the LM, especially in situations where extraneous factors are introduced. Potential ways to mitigate this effect could be refining prompt designs, adjusting model parameters, or incorporating additional validation steps in the AI-assisted risk assessment process. Overall, this study showcases the practical significance of collaborative human and AI analysis in construction industry risk assessments. The findings lay the groundwork for refining methodologies that optimize both human judgment and machine capabilities, offering valuable insights for industries adopting AI in risk assessment. Theoretical contributions stem from the study's categorization of Type I and Type II errors, providing insights into human-AI collaboration dynamics across industries. Future research endeavors will concentrate on enhancing collaborative models, broadening the study's scope, and delving into prompt engineering techniques to refine prompt formulation, integral to advancing this collaborative process.

References

- [1] M. M. Zaira and B. H. W. Hadikusumo, "Structural equation model of integrated safety intervention practices affecting the safety behaviour of workers in the construction industry," 2017, doi: 10.1016/j.ssci.2017.06.007.
- [2] "Fatal occupational injuries by industry and event or exposure, all United States, 2020 : U.S. Bureau of Labor Statistics." Accessed: Nov. 04, 2023. [Online]. Available: <https://www.bls.gov/iif/fatal-injuries-tables/fatal-occupational-injuries-table-a-1-2020.htm>
- [3] "CPWR | A world leader in construction safety and health research and training." Accessed: Nov. 21, 2023. [Online]. Available: <https://www.cpw.com/>
- [4] A. Chokor, H. Naganathan, W. K. Chong, and M. El Asmar, "Analyzing Arizona OSHA Injury Reports Using Unsupervised Machine Learning," *Procedia Eng*, vol. 145, pp. 1588–1593, Jan. 2016, doi: 10.1016/J.PROENG.2016.04.200.
- [5] B. Esmaili and M. Hollowell, "Attribute-Based Risk Model for Measuring Safety Risk of Struck-By Accidents," *Construction Research Congress 2012: Construction Challenges in a Flat World, Proceedings of the 2012 Construction Research Congress*, pp. 289–298, 2012, doi: 10.1061/9780784412329.030.
- [6] M. Zeynalian, B. Trigunarsyah, and H. R. Ronagh, "Modification of Advanced Programmatic Risk Analysis and Management Model for the Whole Project Life Cycle's Risks," *J Constr Eng Manag*, vol. 139, no. 1, pp. 51–59, Apr. 2012, doi: 10.1061/(ASCE)CO.1943-7862.0000571.
- [7] A. Vaswani *et al.*, "Attention Is All You Need," 2017.
- [8] M. R. Hollowell and J. A. Gambatese, "Activity-Based Safety Risk Quantification for Concrete Formwork Construction," *J Constr Eng Manag*, vol. 135, no. 10, pp. 990–998, Apr. 2009, doi: 10.1061/(ASCE)CO.1943-7862.0000071.
- [9] B. Esmaili and M. Hollowell, "Attribute-Based Risk Model for Measuring Safety Risk of Struck-By Accidents," *Construction Research Congress 2012: Construction Challenges in a Flat World, Proceedings of the 2012 Construction Research Congress*, pp. 289–298, 2012, doi: 10.1061/9780784412329.030.
- [10] A. J. P. Tixier, M. R. Hollowell, B. Rajagopalan, and D. Bowman, "Automated content analysis for construction safety: A natural language

- processing system to extract precursors and outcomes from unstructured injury reports,” *Autom Constr*, vol. 62, pp. 45–56, Feb. 2016, doi: 10.1016/J.AUTCON.2015.11.001.
- [11] A. J. P. Tixier, M. R. Hallowell, B. Rajagopalan, and D. Bowman, “Application of machine learning to construction injury prediction,” *Autom Constr*, vol. 69, pp. 102–114, Sep. 2016, doi: 10.1016/J.AUTCON.2016.05.016.
- [12] C. Arteaga, A. Paz, and J. W. Park, “Injury severity on traffic crashes: A text mining with an interpretable machine-learning approach,” *Saf Sci*, vol. 132, Dec. 2020, doi: 10.1016/j.ssci.2020.104988.
- [13] S. A. Prieto, E. T. Mengiste, and B. García de Soto, “Investigating the Use of ChatGPT for the Scheduling of Construction Projects,” *Buildings*, vol. 13, no. 4, Apr. 2023, doi: 10.3390/buildings13040857.
- [14] T. Schomacker and M. Tropmann-Frick, “Language representation models: An overview,” *Entropy*, vol. 23, no. 11, Nov. 2021, doi: 10.3390/e23111422.
- [15] T. B. Brown *et al.*, “Language Models are Few-Shot Learners,” 2020. [Online]. Available: <https://commoncrawl.org/the-data/>
- [16] Kimberly A. Neuendorf, *The content analysis guidebook*. Thousand Oaks, CA: Sage, 2002.
- [17] Klaus H. Krippendorff, *Content analysis: An introduction to its methodology*. Thousand Oaks, CA: Sage, 2004.