# Generative AI for Construction Safety: Fine-Tuning LLaMA-2 With OSHA Injury Reports and Retrieval-Augmented Generation

Rizwan U. Farooqui[1], Abid Hassan[1], Asad Waqar Malik[1]

[1] Mississippi State University, Starkville, MS 39762, USA
rfarooqui@caad.msstate.edu

**Abstract**
Construction workers often face delays in getting clear answers to safety questions, and the lack of timely support can lead to mistakes and preventable incidents. A generative AI assistant can provide workers with instant, context-aware guidance when they cannot reach a safety officer or hesitate to seek help.This paper presents an early prototype developed as the first step in a larger program on AI-driven construction safety. We built a dataset from Occupational Safety and Health Administration (OSHA) severe injury reports by converting incident narratives into question–answer pairs covering event type, nature of injury, body part affected, and source of injury. Using Quantized Low-Rank Adaptation (QLoRA), we fine-tuned the LLaMA-2 7B model with 4-bit nf4 quantized weights and 16-bit (bfloat16) compute to train the system for narrative classification, report summarization, compliance-style Q&A, and drafting safety reports. To enhance reliability, we integrated a retrieval-augmented generation (RAG) pipeline that grounds responses in OSHA standards and similar past cases. Preliminary evaluation indicates that the fine-tuned model shows promising performance, particularly for injury nature and body part classification tasks, establishing a foundation for future improvements. Applications include post-incident analysis, compliance guidance, and safety training support. This baseline study demonstrates the feasibility of a generative AI assistant for construction safety and lays the groundwork for a comprehensive system that integrates safety guidelines, plans, and records to improve safety practices on construction sites.

**Keywords**
Construction safety, Generative AI, Large language models, Retrieval-augmented generation, OSHA injury reports.

## 1. Introduction
The construction industry is consistently recognized as one of the most hazardous work environments, with high rates of accidents and injuries compared to other sectors (Bureau of Labor Statistics, 2024). Despite decades of regulation, many incidents remain preventable. Research shows that safety risks are often tied not only to the dangerous nature of construction tasks but also to gaps in awareness and communication between workers and safety supervisors (Giri, 2020; Pinto et al., 2011). Workers may lack timely access to clear guidance, and safety protocols are frequently perceived as complex or impractical during daily operations. These challenges have contributed to the persistence of unsafe practices and recurring incidents on worksites.

### 1.1 AI and Machine Learning in Construction Safety
Recent advances in artificial intelligence (AI) and machine learning (ML) have created new opportunities to improve construction safety. Applications of AI have been explored across project management and lifecycle activities, showing benefits for efficiency, risk detection, and sustainability (Kazeem et al., 2023; Datta et al., 2024). In the safety domain, ML models have been applied to classify worker-reported incidents and automate analysis of safety reports, offering accuracy beyond manual methods (Bugalia et al., 2022).

Computer vision approaches, in particular, have proven effective in detecting unsafe behaviors and mapping them to relevant safety rules, improving hazard recognition at scale (Fang et al., 2023). While these studies demonstrate technical potential, most solutions focus on narrow tasks, lack transparency, and have not been developed as comprehensive onsite safety assistants that can actively support workers in real time with guidance and documentation.

### 1.2 Generative AI and Conversational Safety Tools

Generative AI, especially large language models (LLMs), has expanded the possibilities for interactive safety support. Early research suggests that conversational assistants can help workers better understand protocols and improve training outcomes by delivering information in natural language (Foosherian et al., 2023; Zheng et al., 2023). However, existing studies rarely adapt models to construction-specific contexts or align them with regulatory frameworks such as OSHA standards. Without domain-specific fine-tuning, generic LLMs may produce inaccurate or incomplete guidance, limiting their usefulness on construction sites.

### 1.3 Research Gap and Contribution

Although AI has been used for predictive modeling, hazard detection, and automated safety reporting, little work has explored the development of a generative AI assistant trained directly on construction safety incidents and regulations. To address this gap, this study presents a fine-tuned LLaMA-2 model integrated with retrieval-augmented generation (RAG). The system is trained on OSHA severe injury reports reformulated into safety-focused question–answer pairs. It is designed to perform multiple tasks, including classification of incidents, summarization of reports, compliance-oriented Q&A, and interactive safety training. By documenting this groundwork, the paper contributes a reproducible foundation for future efforts to build domain-specific generative AI assistants that support safety decision-making and move the industry toward a zero-accident vision. This work should be understood as baseline research that lays the foundation for the larger on site generative artificial intelligence-based construction assistant project.

## 2. Materials and Methods.

This study describes the development of retrieval-augmented generative AI assistant for construction safety. The workflow involved four stages: (1) data collection and preprocessing, (2) model fine-tuning, (3) integration of retrieval-augmented generation (RAG) and (4) evaluation. Figure 1. Overview of the research methodology showing the four main stages below

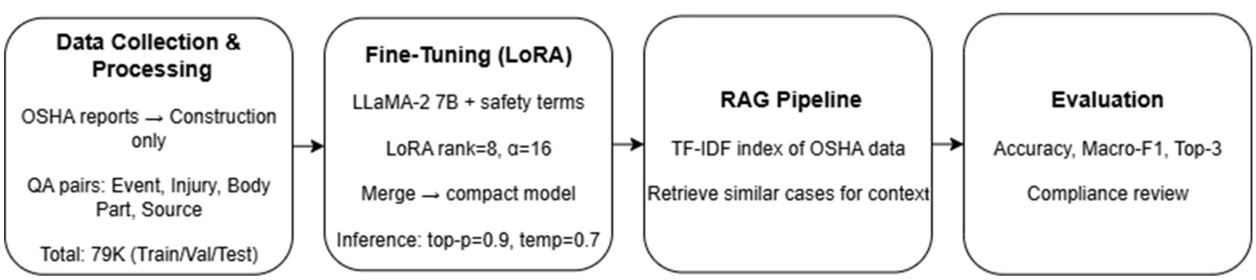


**Figure 1.** Overview diagram of the research methodology

### 2.1 Data Collection and Preprocessing

The primary data source was the Occupational Safety and Health Administration (OSHA) Severe Injury Reports, containing 95,979 total records. To focus on construction-related incidents, the dataset was filtered using the NAICS code, where records with codes beginning with 23 were retained. After filtering, 17,158 records remained.

These reports include narrative descriptions of workplace accidents, providing details such as the type of incident, body part affected, and causal factors. This construction-specific subset was selected to ensure relevance, following earlier studies that emphasize the importance of domain-specific safety datasets (Pinto et al., 2011; Giri, 2020).

Each narrative was then reformulated into multiple supervised question–answer (QA) pairs corresponding to OSHA-style classification labels. Specifically, we generated QA pairs for event or exposure, nature of injury, body part affected, and source of injury. On average, each incident narrative generated approximately 4–5 QA pairs, resulting in a total of 79,035 QA pairs. The dataset generated QA pairs on below categories: (i) Event or exposure; (ii) Nature of injury; (iii) Body part affected; and (iv) Source of injury.

This process transformed each incident narrative into several short input–output examples suitable for instruction-style fine-tuning. After preprocessing, the dataset contained:

- **Train:** 74,035 QA pairs
- **Validation:** 2,500 QA pairs
- **Test:** 2,500 QA pairs
- **Total:** 79,035 QA pairs

This structure was designed to align with OSHA reporting requirements and with practical construction safety tasks. Such reformulation has been shown to support classification, compliance tracking, and safety training (Bugalia et al., 2022; Kazeem et al., 2023).

### 2.2 Model Fine-Tuning

The foundation model was LLaMA-2 with 7 billion parameters, accessed through the Hugging Face Transformers library. To adapt the model without prohibitive resource demands, we used Quantized Low-Rank Adaptation (QLoRA), an extension of LoRA designed for resource-efficient fine-tuning. QLoRA combines 4-bit quantization of the base model weights with 16-bit (bfloat16) compute precision, significantly reducing GPU memory usage while maintaining model quality. This approach enables targeted adaptation with less than 1% of the model's parameters being trainable, striking a balance between efficiency and expressiveness (Datta et al., 2024).

#### 2.2.1 Tokenizer and Input Processing

The tokenizer was extended with construction-specific terms (e.g., "scaffolding," "trench collapse," "OSHA 1926") to reduce segmentation errors and improve domain coverage. QA pairs were pre-tokenized with a maximum sequence length of 512 tokens, with dynamic padding and truncation applied to optimize GPU utilization during training.

#### 2.2.2 LoRA Configuration

LoRA adapters were inserted into the attention projection layers (q_proj and v_proj) of the transformer. The adapters were configured with a rank ($r$) of 8, a scaling factor (α) of 16, and a dropout rate of 0.05. The base model weights were quantized using the nf4 (normalized float 4) data type, which supports effective 4-bit precision while minimizing information loss. This configuration allowed the model to be fine-tuned efficiently without exceeding hardware constraints.

#### 2.2.3 Training Setup

Training was performed on a single NVIDIA A100 GPU (40 GB memory) in a Google Colab Pro environment. The batch size was 8, the learning rate was 2e-5, and weight decay was set to 0.01. Mixed precision with bfloat16 compute was employed for stability and speed, while gradient checkpointing further reduced memory usage.

At the end of training, the LoRA adapter weights were saved and merged into the base model, resulting in a compact, domain-adapted model that no longer required LoRA-specific runtime dependencies. During inference, the model used nucleus sampling (top-p = 0.9) with a temperature of 0.7 and a maximum output length of 100 tokens.

#### 2.2.4 Justification of Approach

Full fine-tuning of LLaMA-2 7B was avoided due to the prohibitive GPU memory requirements (>200 GB). LoRA was selected as enabled targeted adaptation with a small fraction of trainable parameters while preserving the general linguistic capabilities of the base model. This approach provided a practical balance of efficiency, domain adaptation, and reproducibility.

### 2.3 Retrieval-Augmented Generation Integration

While fine-tuned models improve domain knowledge, generative models may still produce incomplete or inaccurate responses. To address this, a retrieval-augmented generation (RAG) pipeline was added, drawing on OSHA regulatory text and similar past incidents. A TF-IDF index was built from training narratives and paired labels, and a nearest-neighbor search was used to retrieve context during inference. Similar designs have been explored in conversational agents to enhance factual accuracy and user trust (Foosherian et al., 2023; Zheng et al., 2023).

### 2.4 Evaluation

Evaluation was performed using both automated metrics and task-specific accuracy measures. For classification tasks, strict accuracy, macro-F1, and top-3 accuracy was calculated. Parent-level accuracy was also reported for complex categories such as incident type and source. These metrics are commonly used in construction safety analytics (Fang et al., 2023; Bugalia et al., 2022).

In addition, qualitative inspection of generated responses was carried out to assess clarity and compliance grounding. While limited in scope, this step provided practical insights into how well the system supports safety-related decision-making and training tasks.

## 3. Results

The fine-tuned model was evaluated on four core classification tasks: event/exposure, nature of injury, body part affected, and source of injury. Performance was measured using strict accuracy, macro-F1, top-3 accuracy, and parent-level accuracy where applicable. These metrics provided complementary views of how well the model captured both fine-grained and higher-level safety labels.

### 3.1 Quantitative Results

Table 1 presents the classification results.

**Table 1**. Classification results for construction safety tasks

| Task | Strict Accuracy | Macro-F1 | Top 3 Accuracy | Parent-Level Accuracy |
|---|---|---|---|---|
| Event/Exposure | 21.71% | 0.1954 | 44.90% | 48.19% |
| Nature of Injury | 66.22% | 0.4992 | 79.85% | Not Applicable |
| Body Part Affected | 71.23% | 0.5599 | 82.87% | Not Applicable |
| Source of Injury | 17.27% | 0.1649 | 35.72% | 40.24% |

Figure 2 provides a visual comparison of strict accuracy, macro-F1, top-3 accuracy, and parent-level accuracy. The chart highlights the strong performance on nature and body part tasks and shows the relative gap in event and source classification. Figure 3 compares fine-grained and parent-level performance for event and source. The figure illustrates that while fine-grained prediction is challenging, broader hazard patterns are captured more effectively, with improvements of over 20 percentage points at the parent level.

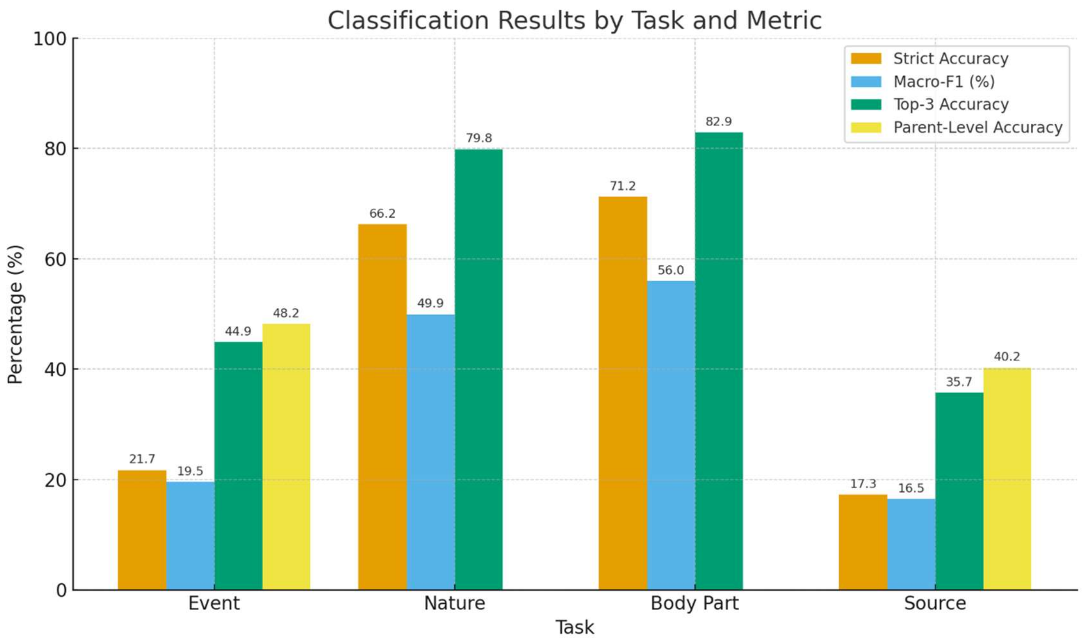


**Figure 2**. Classification performance of the fine-tuned model

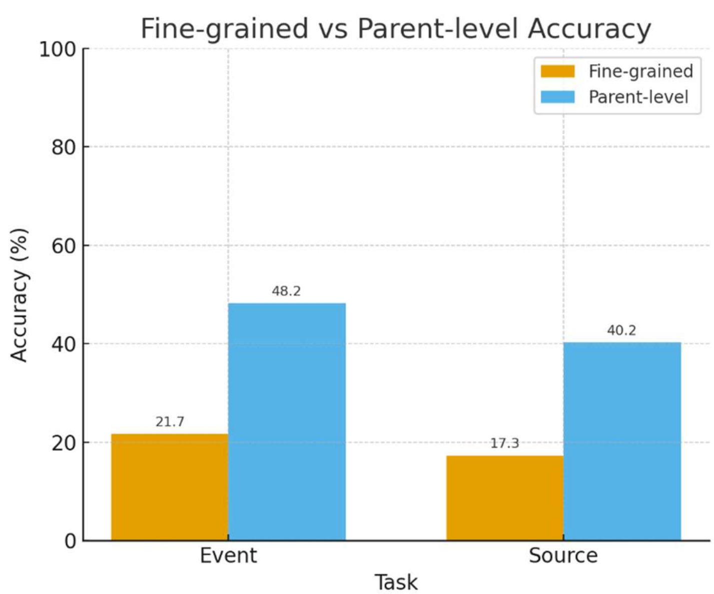


**Figure 3.** Comparison of fine-grained versus parent-level accuracy for event and source classification tasks.

## 4. Discussion

### 4.1 Interpretation of Results

The results show that fine-tuning a large language model on OSHA narratives can produce useful outputs, but performance is uneven. The model achieved higher accuracy on nature of injury and body part affected, both above 65%. These categories are more consistent, with clearer wording in the reports, which made them easier for the model to learn. By contrast, performance was much lower on event and source classification. These categories are harder because they contain many overlapping or rarely used labels. For example, “fall from ladder” and “fall from roof” often appear in similar contexts, which confuses the model. Parent-level evaluation confirmed that the model can capture broad hazard types even when it fails on fine-grained labels. This shows that the system has value for identifying general patterns but is not yet ready for reliable detailed classification.

### 4.2 Comparison with Prior Work

The performance trends observed here are consistent with earlier construction safety studies, where machine learning methods performed better on structured tasks and struggled with diverse or noisy categories (Bugalia et al., 2022; Fang et al., 2023). At the same time, this work connects to broader LLM literature. Many recent applications of large language models have concentrated on domains such as mental health and counseling, where conversational agents are trained on curated datasets and tasks are narrower and more clearly defined (Jin et al., 2025; Hua et al., 2025; Cho et al., 2023). These contexts provide abundant data and relatively low label ambiguity, which makes adaptation more straightforward. By contrast, construction safety presents a noisier, under-resourced environment with overlapping categories and inconsistent reporting. This difference helps explain why performance on event and source classification remains lower, and why LLM-based safety assistants remain largely unexplored. In this sense, work contributes baseline evidence in a domain where generative AI has received very limited attention.

### 4.3 Strengths and Contributions

This work makes several contributions. It demonstrates that parameter-efficient fine-tuning with LoRA can adapt a general model to a specialized safety domain using modest hardware. It also shows that one model can support multiple safety tasks, not just a single classification job. The integration of retrieval-augmented generation helped ground answers in OSHA standards, which reduced unsupported outputs and made compliance-related answers more trustworthy. Together, these elements form a baseline that other researchers can reproduce, extend, and compare against. The value of this study lies less in the raw scores and more in showing what can be achieved, what cannot, and what should be improved.

### 4.4 Limitations and Critical Reflection

This study has several limitations. The model was trained on 74,000 QA pairs, which, while substantial, represents only a portion of the potential data available from OSHA and other sources. A larger and more diverse dataset could significantly improve performance, particularly for challenging categories like event and source. In addition, the OSHA label space is inconsistent, with overlapping and imbalanced classes, making fine-grained classification difficult without further label cleaning and consolidation.

The LoRA fine-tuning parameters were chosen based on initial experimentation and may not be optimal. Future research should systematically explore alternative configurations to improve domain adaptation. Resource and training time constraints also limited experimentation; only a single GPU was used, and techniques such as model distillation or multi-stage fine-tuning were not implemented, though they could enhance scalability and deployment readiness.

While a RAG pipeline was integrated to improve factual grounding, its performance was not separately evaluated, so its contribution remains unclear. Similarly, the question-setting patterns used to transform OSHA narratives into QA pairs were not systematically studied, and refining these could improve model understanding of safety scenarios. Finally, the evaluation relied mainly on basic metrics like accuracy and F1-score, which do not fully measure the model's ability to generate summaries or compliance guidance. Future studies should expand evaluation methods and explore advanced retrieval techniques, such as dense or hybrid search, to improve overall performance and reliability.

**4.5 Pathway for Future Work**

This work should be understood as baseline research. The purpose was not to deliver a production-ready system but to explore feasibility and document both strengths and weaknesses. The analysis shows clear pathways for improvement: training on the full dataset, consolidating and cleaning labels, adding more advanced retrieval, and broadening the evaluation framework. Future studies may also expand beyond text by integrating other domain resources such as construction safety manuals or site-specific documentation. By presenting both the successes and the shortcomings, this study provides a foundation that others can build on. It marks an exploratory step toward creating generative AI systems that can eventually support construction workers with reliable, real-time safety guidance. Future extensions will be built on this baseline to expand the project into a more comprehensive system that integrates additional data sources and safety documentation.

**5. Conclusions**

This study presented a baseline for generative AI assistant for construction safety, developed by fine-tuning LLaMA-2 7B on OSHA severe injury narratives and integrating a retrieval-augmented generation pipeline. The model supported multiple tasks, including classification, summarization, and compliance-oriented Q&A. Results showed that the system performed well on structured categories such as nature of injury and body part affected, while more diverse categories such as event and source remained challenging. Parent-level evaluation confirmed that broader hazard patterns could still be captured, even when fine-grained labels were missed.

The contribution of this work lies not in achieving perfect accuracy but in establishing a reproducible benchmark and analyzing its strengths and weaknesses. By identifying where the model succeeds and where it falls short, this study provides a pathway for future research. Improvements in dataset coverage, label consolidation, evaluation methods, and retrieval strategies are expected to strengthen performance, and the integration of additional resources such as safety manuals or site documentation could further enhance practical utility. Overall, this work should be understood as an exploratory step that demonstrates the feasibility of applying large language models to construction safety and sets the stage for future systems that may one day provide reliable, real-time guidance to workers and supervisors.

From a practical perspective, this baseline shows that generative AI can already help with tasks such as incident summarization, linking cases to OSHA standards, and providing structured feedback for training. While not ready for deployment, the framework highlights how future systems could act as instant, context-aware assistants for workers and supervisors who may not always have access to safety experts on site. Overall, this study should be seen as a baseline contribution to the larger research initiative, which seeks to develop reliable, real-time generative AI assistance for construction safety.

**References**

Bureau of Labor Statistics. (2024, May 30). A look at falls, slips, and trips in the construction industry. U.S. Department of Labor. https://www.bls.gov/opub/ted/2024/a-look-at-falls-slips-and-trips-in-the-construction-industry.htm

Bugalia, N., Tarani, V., Kedia, J., & Gadekar, H. (2022). Machine learning-based automated classification of worker-reported safety reports in construction. Journal of Information Technology in Construction, 27, 926–950. https://doi.org/10.36680/j.itcon.2022.045

Cho, Y. M., Rai, S., Ungar, L., Sedoc, J., & Guntuku, S. C. (2023). An integrative survey on mental health conversational agents to bridge computer science and medical perspectives. arXiv Preprint, arXiv:2310.17017. https://arxiv.org/abs/2310.17017

Datta, S. D., Islam, M., Rahman Sobuz, M. H., Ahmed, S., & Kar, M. (2024). Artificial intelligence and machine learning applications in the project lifecycle of the construction industry: A comprehensive review. Heliyon, 10(5), e26888. https://doi.org/10.1016/j.heliyon.2024.e26888

Fang, W., Love, P. E. D., Ding, L., Xu, S., Kong, T., & Li, H. (2023). Computer vision and deep learning to manage safety in construction: Matching images of unsafe behavior and semantic rules. IEEE Transactions on Engineering Management, 70(12), 4120–4132. https://doi.org/10.1109/TEM.2021.3093166

Foosherian, M., Purwins, H., Rathnayake, P., Alam, T., Teimao, R., & Thoben, K.-D. (2023, September). Enhancing pipeline-based conversational agents with large language models. In D. Hazarika, X. R. Tang, & D. Jin (Eds.), Proceedings of the 1st Workshop on Taming Large Language Models: Controllability in the Era of Interactive Assistants! (pp. 56–67). Association for Computational Linguistics. https://aclanthology.org/2023.tllm-1.6/

Giri, O. P. (2020). Factors causing health and safety hazards at construction sites. Technical Journal, 2(1), 68–74. https://doi.org/10.3126/tj.v2i1.32841

Hua, Y., Na, H., Li, Z., Liu, F., Fang, X., Clifton, D., & Torous, J. (2025). A scoping review of large language models for generative tasks in mental health care. NPJ Digital Medicine, 8, 230. https://doi.org/10.1038/s41746-025-01611-4

Jin, Y., Liu, J., Li, P., Wang, B., Yan, Y., Zhang, H., Ni, C., Wang, J., Li, Y., Bu, Y., & Wang, Y. (2025). The applications of large language models in mental health: Scoping review. Journal of Medical Internet Research, 27, e69284. https://doi.org/10.2196/69284

Kazeem, K. O., Olawumi, T. O., & Osunsanmi, T. (2023). Roles of artificial intelligence and machine learning in enhancing construction processes and sustainable communities. Buildings, 13(8), Article 2061. https://doi.org/10.3390/buildings13082061

Pinto, A., Nunes, I. L., & Ribeiro, R. A. (2011). Occupational risk assessment in construction industry: Overview and reflection. Safety Science, 49(5), 616–624. https://doi.org/10.1016/j.ssci.2011.01.003

Zheng, Z., Liao, L., Deng, Y., & Nie, L. (2023). Building emotional support chatbots in the era of LLMs. arXiv Preprint, arXiv:2308.11584. https://doi.org/10.48550/arXiv.2308.11584