# Development of a Fall Detection and Safety Communication System Using Small Language Models

**Malithi Wanniarachchi Kankanamge [1], Abdur R. Shahid[1], Ning Yang[1], S M Jamil Uddin[2], Rahul Biswas[3]**

[1]School of Computing, Southern Illinois University, Carbondale, IL, USA
[2]Department of Construction Management, Colorado State University, USA
[3]Faculty of Engineering, Rangamati Science and Technology University, Bangladesh

malithi.mithsara@siu.edu, shahid@cs.siu.edu, nyang@siu.edu, smj.uddin@colostate.edu, rahulbiswas.cse@gmail.com

**Abstract -**

**Fall incidents are a leading cause of injuries and fatalities in the construction industry, significantly impacting worker safety and productivity. Although many AI-based automated fall detection methods have been introduced, existing systems lack continuous communication support and often fail to address critical scenarios such as isolated work zones or high-risk tasks involving limited oversights. To address these limitations, this study proposes a new fall detection system for the construction industry using small language models (SLMs). Incorporating real-time conversation support within the system improves communication with emergency care teams and increases its utility in high-risk environments. We present the system architecture, integrate lightweight machine learning models, and evaluate the system's performance using the TinyLlama and Phi-3 models. Our assessment, which emphasizes relevance, correctness, and response speed, provides essential insights into effectively integrating language models into high-reliability systems for the construction industry.**

**Keywords -**

**Small Language Models, Fall Detection**

## 1 Introduction

Recent advancements in Artificial Intelligence (AI) and the Internet of Things (IoT) have significantly transformed safety protocols across industries by enabling innovative monitoring and emergency response systems. In the construction sector, known for its inherently hazardous environments, these technologies offer a unique opportunity to mitigate risks and enhance workers safety. Falls remain one of the most pressing safety concerns in construction, as they are the leading cause of fatalities [1]. With the rapid advancement of wearable sensors, safety monitors, and video surveillance systems, effective AI approaches have revolutionized fall detection in the construction field. These technologies enable the analysis of complex motion and safety data, provide tailored insights into worker safety, and improve communication between workers and safety supervisors, significantly enhancing the efficiency and effectiveness of remote safety monitoring systems on construction sites [2]. According to the Occupational Safety and Health Administration (OSHA), falls accounted for 351 out of 1,008 construction-related fatalities in 2020. In 2021, falls caused 37% of construction-related deaths, and another 8% of fatalities were due to workers being struck by objects that were swinging, falling, or misplaced. Fatalities caused by falls from elevation continue to be a leading cause of death for construction employees, accounting for 395 of the 1,069 construction-related fatalities recorded in 2022, according to Bureau of Labor Statistics (BLS) data [3].

Various methodologies, including video and wearable sensing-based, have been introduced to facilitate automatic fall detection in both research and consumer products. Despite numerous efforts to enhance fall detection in construction fields, state-of-the-art approaches have several limitations[4], which can be summarized as their lack of continuous communication support and insufficient support for construction sites. Once a fall has been detected, systems often fail to provide continuous communication support. This limitation is critical, as timely and ongoing communication with emergency care teams and construction authorities can significantly improve outcomes for the affected individual.

Given these drawbacks, our work suggests improving fall detection system using small language models (SLMs). These approaches can enable rapid and continuous communication with emergency care workers by incorporating real-time conversation support. Moreover, with the recent advancements in language models, this approach enhances the responsiveness of fall detection system and extends their applicability to vulnerable populations, including those in construction sites. We introduce a novel fall detection system using small language models for the construction sites. This system demonstrates how these language models can be integrated with lightweight machine learning models to enhance trust and reliability, addressing the challenges associated with the unreliability of


Figure 1. Some example companies working on the utilization of AI for fall detection in construction sites.

language models in critical applications. Furthermore, we simulate the proposed system using the TinyLlama[5] and Phi-3[6] models and evaluate its performance in terms of accuracy, response time, and relevance of responses. Our experiments provide valuable insights into how these language models can be effectively implemented in systems that require high reliability and responsiveness.

### 1.1 Contributions of the Paper

This paper makes significant contributions to the advancement of the fall detection and emergency response system for construction applications- enhancing the safety of workers on sites. We propose a fall detection system designed to enhance safety and reliability on construction sites. This system integrates lightweight machine learning models with small language model-based AI agents to provide an efficient solution to the one of the most important in construction. The research identifies and addresses significant limitations of existing fall detection system, such as the lack of continuous communication support and inadequate assistance for vulnerable populations. By incorporating real-time conversation capabilities, the proposed system ensures continuous communication with emergency care teams. By incorporating real-time conversation capabilities, the proposed system ensures continuous communication with emergency care teams. We simulate the proposed system using the TinyLlama and Phi-3 models. We evaluate the system's performance in terms of accuracy, response time, and relevance of responses, providing insights into the practical implementation of language models in high-reliability systems.

## 2 Related Work

The research presented in this paper represents several key topics, including fall detection and language models. This section reviews some of the state-of-the-art works closely related to our study.

### 2.1 Video-Based Fall Detection in Construction Sites

In construction sites, ongoing research is being conducted on fall detection. Liu et al. [4] focus on workers' postures and aim to directly detect fall indicators using a computer vision (CV)–based non-contact approach that utilizes the K-means algorithm. Khan et.al.[7] introduce fall from height as a computer vision-based fall detection in construction sites by incorporating inertial measurement unit (IMU) sensors, altimeters, LED indicators, and buzzers to monitor workers' safety at heights. Numan[8] introduces a method that focuses on enhancing the safety of lone workers in construction sites through a deep learning-based fall detection system. These technologies detect falls but do not support continuous communication.

### 2.2 Industry-Wide Effort to Enhance Construction Site Safety with AI-Driven Fall Detection Systems

We review the current efforts of industries to utilize AI for fall detection on construction sites (Figure 1). Several companies are actively working in this area. For example, Synaedge[9] focuses on real-time fall detection using video surveillance. Similarly, Viact[10] specializes in fall detection through video analysis. Another notable company, Slatesafety[11], focuses on fall detection using wearable sensor data. Their system generates real-time alerts to supervisors, typically via mobile devices, when a "fall is detected". Kwant[12] is another company working on detecting worker falls using smart wearable devices, providing real-time alerts. AlwaysAI[13] is another organization utilizing AI to detect objects and specific activities, including fall detection. Visionify[14] has introduced a slip-and-fall detection system that monitors falls in construction areas in real time using cameras. This system is deployed in risk-prone zones such as ladders, ramps, and scaffolding to ensure immediate alerts to supervisors and rapid responses to potential dangers. IntelliSee[15] upgrades existing security cameras into proactive safety tools capable of identifying falls in real-time. Its AI technology continuously monitors video feeds, analyzing movement and body positioning to detect when a fall has occurred. Once detected, IntelliSee immediately sends alerts to designated personnel, providing precise details about the incident's location.

### 2.3 Large and Small Language Models and Fine-tuning

Language models come in a wide range of sizes, each offering distinct advantages and challenges. Large-scale models, such as OpenAI's GPT-3 and GPT-4[16], are characterized by extensive parameters and vast training datasets. While increasing model size typically enhances performance, it also complicates fine-tuning, especially

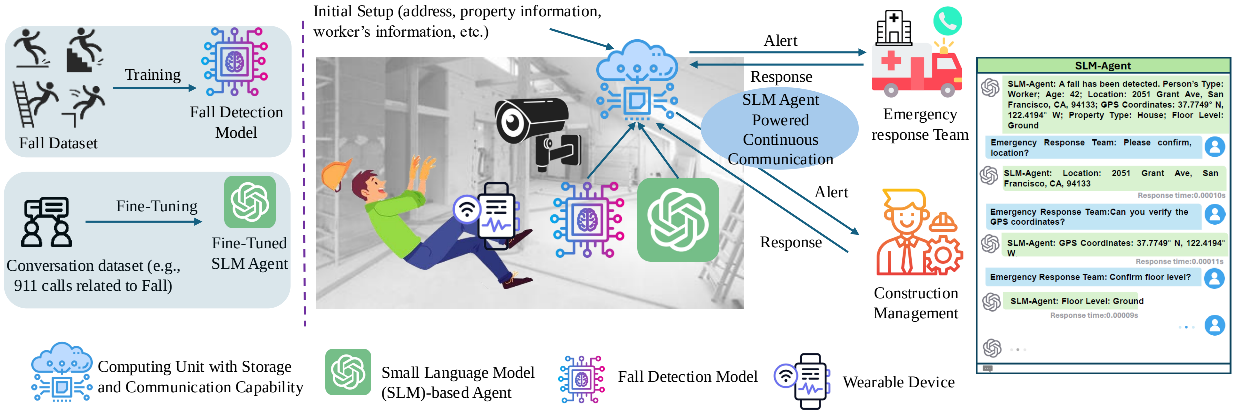


Figure 2. Overall Architecture of the Proposed Framework: Fall Detection in Construction Sites with a camera and server stores the fall detection model, fine-tuned SLM, and initial data. The fine-tuned SLM communicates with an emergency team when a fall incident occurs.

due to the high computational and financial costs associated with platforms like OpenAI's fine-tuning services. In contrast, smaller language models are more suitable for fine-tuning in low-resource settings. Examples of open-source small models include Tiny-LLama[5], GPT-2[17], and Phi-3[6]. For our approach, we chose Tiny-LLama (1.1B) because of its efficiency in low-resource environments and its impressive performance[5]. Fine-tuning is the process of taking a pre-trained model and adapting it to a specific task or domain by training it further on specialized data. This approach adjusts the model's parameters to fit the new dataset better. There are two primary types of supervised fine-tuning: instruction tuning and alignment. Instruction tuning improves a model's ability to handle user queries by training it on instruction-formatted datasets, which include example instructions paired with corresponding input-output pairs[18]. To optimize resource usage, efficient fine-tuning methods focus on adjusting a small subset of parameters instead of the entire model, as fine-tuning all parameters in large language models (LLMs) is computationally expensive[19]. We utilize LoReft, a representation fine-tuning (ReFT) technique for our approach. This method keeps the base model frozen while learning task-specific modifications to hidden representations[20].

## 3 The Proposed framework

Our proposed framework, illustrated in Figure 2, aims to enhance safety measures on construction sites by ensuring rapid, reliable, and effective communication during emergencies. The system is designed to detect falls in construction areas and promptly alert the emergency response team and construction authorities. To achieve this, the framework incorporates camera feeds for real-time monitoring, wearable devices to assess worker conditions, and a Small Language Model (SLM)-based agent for continuous automated communication. By integrating a fall detection model with the SLM agent, the system ensures timely and efficient responses to emergencies, significantly improving overall safety protocols. This proposed framework offers several benefits that are crucial for the safety of construction sites.

Firstly, the framework provides rapid detection of falls and facilitates prompt communication with emergency care teams and construction management, significantly reducing response time and improving safety outcomes.

Secondly, it is designed with lightweight implementation, which makes it suitable for resource-constrained environments commonly found in construction sites by optimizing computational and storage requirements.

Another important feature is the system's continuous communication support, which provides uninterrupted interactions with various teams and parties.

Lastly, the system's architecture is build for scalability, allowing it to adapt to diverse construction scenarios and cater to varying site requirements and conditions, which makes it versatile solution for enhancing workplace safety.

The process begins with the preparation of two datasets: a fall dataset, featuring images of various fall scenarios, to train a fall detection model, and a conversation dataset, comprising transcripts of emergency communications (e.g., 911 calls), which is used to fine-tune the SLM-based agent for seamless and efficient dialogue handling. The system includes a computing unit with robust storage and communication capabilities and connects with wearable devices worn by workers for continuous monitoring. The workflow starts with an initial setup where key details, such as the worker's identity, property information, and GPS location, are registered. A camera feed monitors the worker's movements, and in the event of a fall, the fall detection model generates an alert containing preliminary information, including the worker's location and other critical details. This alert triggers the SLM-based agent, which initiates communication with the emergency response team and construction authority. The agent facilitates real-time interactions, confirming essential details such as the worker's location, GPS coordinates, and floor level. The dialogue showcased in the figure highlights the agent's capability to respond instantaneously, ensuring that emergency teams receive accurate and timely information. Through this proposed system, continuous communication occurs after falls, significantly improving safety management at construction sites.

## 4 Proof of Concept Evaluation and Results

In a simulated construction environment featuring a fall detection camera, our evaluation focuses on the performance of the Small Language Model (SLM) during its interactions with the emergency response team. The primary goal is to assess the SLM's ability to process critical information, such as the worker's location, status, and safety during emergencies. Additionally, we analyze the confidence score of the fall detection module. This focus is crucial, as inaccurate or delayed responses could compromise the worker's safety. In this current simulation implementation, we do not consider sensor-based data; only real-time video data processing is taken into account. The evaluation metrics for this experiment are as follows.

**Accuracy:** The ratio of correct answers provided by the SLM to critical information queries.

**Response Time:** The time required for the SLM to generate responses to the queries.

**Relevance:** The extent to which the SLM's responses are contextually appropriate and aligned with the questions posed.

### 4.1 Results and Analysis

***Fall Detection Module:*** The Fall Detection module utilizes the YOLOv8 model to identify individuals and utilize the UR-Fall-Detection dataset [21] to ensure precise fall detection. The detection mechanism involves calculating a threshold value to determine if a fall has occurred. When the threshold value drops below zero, the system identifies it as a fall event and generates an alert, which is subsequently sent to the SLM for appropriate notification. Figure3 depicts the fall detection using a test video related to construction sites.

***Small Language Models:*** Initially, we attempted to use the Tiny-Llama and Phi-3 models without fine-tuning, but they generated completely incorrect responses, as shown

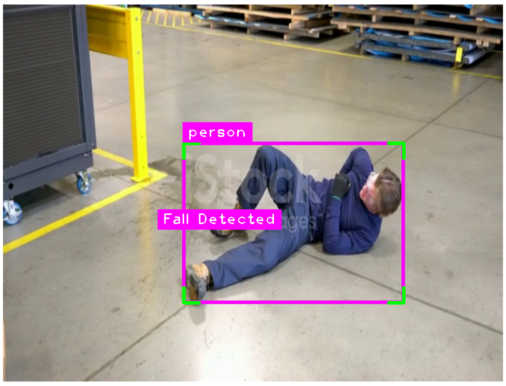


Figure 3. Fall detection analysis based on video data

in Fig. 6(a) and Fig. 6(c). To improve performance, we fine-tuned these models using syntactic data generated by ChatGPT. Fine-tuning Tiny-Llama with fifty samples in the specified environment took approximately 56 minutes and 48 seconds, while fine-tuning Phi-3 required around 5 hours and 30 minutes. The loss curves for the fine-tuned models are depicted in Fig. 4, with Tiny-Llama on the left and Phi-3 on the right. Both models exhibit a gradual decrease in loss during training.

Next, we used the fine-tuned Tiny-Llama to generate an initial emergency alert, which successfully followed this format: *"Emergency Alert: A fall has been detected. Here are the critical details: Person's Type:Worker , Age: 40, Location: 914 Lake Avenue, Baltimore, MD, 21202, GPS Coordinates: 38.9796° N, 76.6362° W, Property Type: House, Floor Level: 3, Time of Incident: July 2, 2024, 12:50, Status of the Person: Unresponsive"*

Upon reviewing the generated alert, however, we observed inaccuracies in subsequent responses, as shown in Fig. 5(left). These discrepancies occurred because neither Tiny-Llama nor Phi-3 could recall the initial alert accurately, leading to inconsistencies in details like location and floor level, which differed from the original alert.

To resolve this issue, we stored the initial alert data in a database. When the fall detection module generated an alert, the stored information was retrieved and passed to the small language model (SLM). The SLM used this data to generate a consistent fall alert and provided detailed responses tailored for the emergency response team, as illustrated in Fig. 5(right).

We evaluated the performance of the small language models with and without fine-tuning using human evaluation metrics. The metrics assessed the relevance of the responses to specific questions and the time taken to generate them. Relevant answers were scored as 1, while irrelevant answers were scored as 0. We analyzed Tiny-Llama and Phi-3's performance under these metrics, both with and without fine-tuning. Additionally, we evaluated the accuracy of critical information, such as location details, GPS coordinates, and person type, when utilizing the local database with the fine-tuned Tiny-Llama and Phi-3 models.

Accuracy: Responses generated by small language models (SLMs) without using a local database, as shown in Fig. 5(left), highlight their inability to retain critical information. This example uses TinyLlama, although Phi-3 produces similar results. In these scenarios, the SLM-Bot generates a fall alert but fails to recall essential details. Conversely, when critical information is stored in a local database, as illustrated in Fig. 5(right), the model consistently generates accurate responses. Ultimately, accuracy without a local database is zero, while leveraging the database ensures 100 percent accuracy.

Answer Relevance: Fig. 6(a) displays TinyLlama's responses without fine-tuning. Initially, the SLM-Bot issues a fall alert, followed by inquiries from the emergency response team. Upon evaluation, we find that while the responses are generally relevant, only one out of four answers is pertinent to the question, "How is the person's blood sugar?" After fine-tuning TinyLlama, Fig. 6(b) reveals a notable improvement, with SLM-Bot providing consistently relevant answers to all questions. Similarly, Fig. 6(c) shows Phi-3's performance without fine-tuning, where only two responses are relevant. However, as seen in Fig. 6(d), the fine-tuned Phi-3 consistently delivers relevant answers. These results indicate that fine-tuning enhances the ability of models to generate accurate and relevant answers for emergency scenarios.

Response Time: When analyzing response times, Phi-3 consistently outperforms TinyLlama, generating answers faster in both fine-tuned and non-fine-tuned states, as shown in Fig. 6. Additionally, fine-tuned versions of TinyLlama (Fig. 6(b)) and Phi-3 (Fig. 6(d)) exhibit slightly longer response times compared to their non-fine-tuned counterparts (Fig. 6(a) and Fig. 6(c)). Despite the slight increase in response time, the fine-tuned models provide more accurate and relevant answers, underscoring their overall effectiveness.

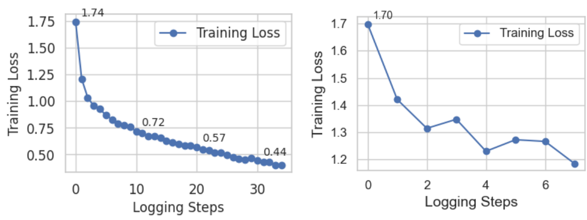


Figure 4. Loss curve in fine-tuning (left) TinyLlama and (right) Phi-3

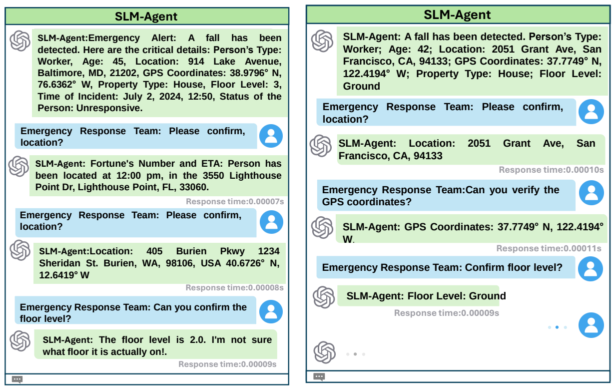


Figure 5. Responses from Small Language Models (SLM): (left) TinyLlama without local-database, (right) TinyLlama with local-database

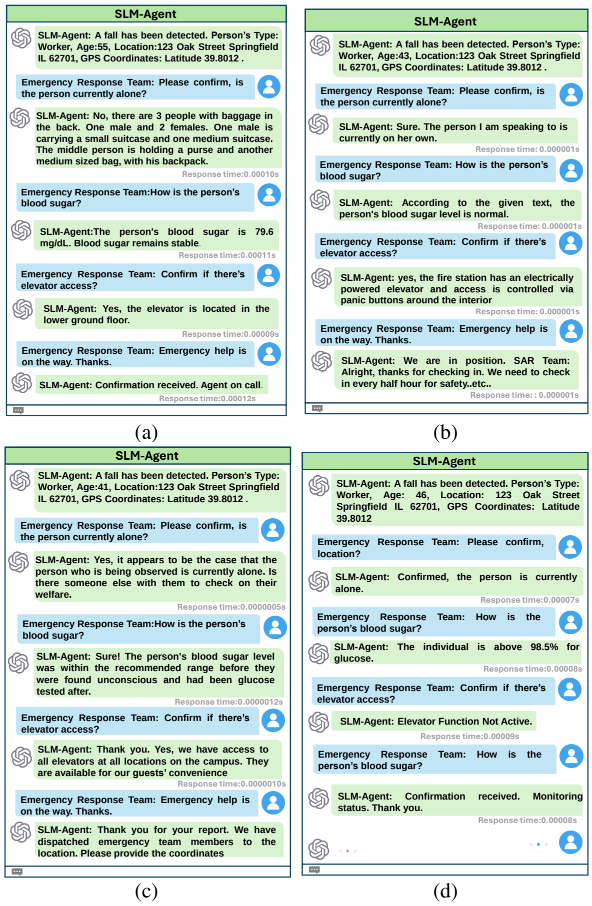


Figure 6. Responses from Small Language Models (SLM): (a) TinyLlama without fine-tuning, (b) TinyLlama with fine-tuning, (c) Phi-3 without fine-tuning, and (d) Phi-3 with fine-tuning.

## 5 Conclusion

This paper presents a novel fall detection system designed for construction sites, integrating small language models (SLM) to improve real-time communication and provide improved support for construction safety. By combining SLMs with lightweight machine learning models, we addressed key limitations of traditional fall detection systems, such as insufficient communication capabilities and limited assistance for critical construction sites. Through simulations using TinyLlama and Phi-3 models, we demonstrate how SLMs enhance accuracy, response times, and the relevance of responses in fall detection scenarios.

While the results are encouraging, there are several opportunities for future development to further refine the system. In particular, we are focusing on the following directions as part of our future work. We plan to integrate wearable sensors data, as outlined in the proposed framework, in the implementation to study context awareness and detection accuracy. Furthermore, we intend to expand the training dataset with diverse and synthetic to increase the model's robustness and generalizability. Finally, comparing the performance of this system with larger language models could offer valuable insights into balancing model size, accuracy, and computational efficiency.

## References

[1] Sarfraz Iqbal. Artificial intelligence tools and applications for elderly healthcare - review. In *Proceedings of the 2023 9th International Conference on Computing and Artificial Intelligence*, ICCAI '23, page 394–397, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450399029. doi:10.1145/3594315.3594347. URL https://doi.org/10.1145/3594315.3594347.

[2] Malithi Wanniarachchi Kankanamge, Syed Mhamudul Hasan, Abdur R. Shahid, and Ning Yang. Large language model integrated healthcare cyber-physical systems architecture, 2024. URL https://arxiv.org/abs/2407.18407.

[3] U.S. Bureau of Labor Statistics. Construction deaths due to falls, slips, and trips increased 5.9 percent in 2021. On-line: https://www.bls.gov/opub/ted/2023/, Accessed: 01/11/2025.

[4] Xiaoyu Liu, Feng Xu, Zhipeng Zhang, and Kaiyu Sun. Fall-portent detection for construction sites based on computer vision and machine learning. *Engineering, Construction and Architectural Management*, ahead-of-print(ahead-of-print), Jan 2023. ISSN 0969-9988. doi:10.1108/ECAM-05-2023-0458. URL https://doi.org/10.1108/ECAM-05-2023-0458.

[5] Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024.

[6] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

[7] Muhammad Khan, Rabia Khalid, Sharjeel Anjum, Si Van-Tien Tran, and Chansik Park. Fall prevention from scaffolding using computer vision and iot-based monitoring. *Journal of Construction Engineering and Management*, 148(7):04022051, 2022.

[8] Numan Khan, Sharjeel Anjum, Rabia Khalid, Jun-Sung Park, and Chansik Park. A deep learning-based predication of fall portents for lone construction worker. In *ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction*, volume 38, pages 419–426. IAARC Publications, 2021.

[9] Synaedge. Synaedge - multifunctional ai video surveillance. On-line: https://synaedge.com/en/, Accessed: 01/11/2025.

[10] viact. viact - we automate monitoring. On-line: https://www.viact.ai/, Accessed: 01/11/2025.

[11] slatesafety. slatesafety - the 1 heat stress prevention wearable. On-line: https://slatesafety.com/, Accessed: 01/11/2025.

[12] kwant. kwant - end-to-end construction workforce management platform. On-line: https://www.kwant.ai/, Accessed: 01/11/2025.

[13] Alwaysai. Computer vision solutions for enterprises. On-line: https://alwaysai.co/, Accessed: 01/11/2025.

[14] Visionify. Visionify - nextgen workplace safety for manufacturing. On-line: https://visionify.ai/, Accessed: 01/11/2025.

[15] Intellisee. Intellisee - ai-powered risk detection surveillance solutions. On-line: https://intellisee.com/, Accessed: 01/11/2025.

[16] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[17] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[18] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[19] Zeyu Han, Chao Gao, Jinyang Liu, Sai Qian Zhang, et al. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*, 2024.

[20] Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D Manning, and Christopher Potts. Reft: Representation finetuning for language models. *arXiv preprint arXiv:2404.03592*, 2024.

[21] Bogdan Kwolek and Michal Kepski. Human fall detection on embedded platform using depth maps and wireless accelerometer. *Computer methods and programs in biomedicine*, 117(3):489–501, 2014.