Edge LLMs for Real-Time Contextual Understanding with Ground Robots

Tamil Selvan Gurunathan[*1], **Muhammad Shehrose Raza**[*1], **Aswin Kumar Janakiraman**[1]
Md Azim Khan[1], **Biplab Pal**[1], **Aryya Gangopadhyay**[*1]

[1]Center for Real-time Distributed Sensing and Autonomy
College of Engineering and Information Technology
University of Maryland Baltimore County, Baltimore, Maryland, USA
{cr36577,mraza2,aswinkj1,azimkhan22,bpal1,gangopad}@umbc.edu

Abstract

We propose a novel framework that leverages Edge Large Language Models (LLMs) for real-time decision-making and contextual understanding on robotic platforms. By embedding LLMs directly on edge devices, the system enables autonomous operations in zero-visibility environments such as tunnels, adverse weather, or tactical obstructions. The framework integrates multi-modal sensor inputs, including mmWave radar and thermal cameras, and employs pre-trained LLMs fine-tuned for low-latency inference under strict computational constraints. Experiments demonstrate the framework's ability to navigate, detect threats, and prioritize tasks such as medical assistance, achieving high semantic accuracy, and significantly outperforming baseline methods like Few-Shot Learning and Prompt Engineering. Furthermore, the system is scalable to diverse applications, including search and rescue, tactical operations, and multi-robot coordination. This work highlights the transformative potential of Edge LLMs in enabling intelligent, reliable, and autonomous robotic systems for dynamic and resource-constrained environments.

Introduction

Robotic platforms operating in complex, dynamic environments require robust capabilities for real-time perception, reasoning, and decision-making. Traditional approaches rely on centralized computing to process sensor data, but latency, bandwidth limitations, and the need for autonomy in field deployments necessitate edge computing solutions (Chowdhury et al. 2023; Lee et al. 2021). This paper introduces a novel framework that leverages Large Language Models (LLMs) optimized for edge computing to enhance the autonomy and adaptability of robotic platforms in zero-visibility environments.

Edge-based LLMs enable robotic platforms to process multi-modal inputs—thermal images, radar signals, and environmental audio—locally and in real time (Lewis et al. 2020; Yadav et al. 2022). Embedding LLMs directly on robotic platforms allows for complex contextual interpretation, such as identifying threats or injured individuals in adverse conditions, without relying on remote servers. This capability is critical for search and rescue, tactical operations, and disaster response, where traditional optical sensors often fail.

This work focuses on the deployment of Edge LLMs integrated with on-board sensors, including millimeter-wave (mmWave) radar and thermal cameras. The LLM processes contextual data to generate situational insights, classify scenarios, and recommend appropriate actions. Unlike conventional methods that rely solely on raw sensor processing, the proposed system fine-tunes pre-trained LLMs for low-latency inference, achieving high semantic understanding and adaptability to novel environments.

Experiments validate the efficacy of this framework in real-world scenarios. Ground robots equipped with Edge LLMs were tested in zero-visibility conditions, demonstrating autonomous navigation, threat detection, and task prioritization (e.g., medical assistance). The system enhances decision-making while reducing dependency on high-bandwidth communication links, making it a scalable solution for resource-constrained environments.

This paper contributes to Edge AI and autonomous robotics by presenting a framework that integrates LLMs with robotic platforms. By emphasizing real-time decision-making, it advances autonomous, reliable, and versatile robotic systems capable of operating in the most challenging conditions.

The applications of this technology are broad, from search and rescue missions, where precise navigation is crucial, to military operations in close-quarters combat, where visibility is compromised. Our experiments include deploying ground robots like HUSKY and SPOT in smoke-filled tunnels, enabling multi-robot coordination with thermal cameras and gesture or voice guidance for situational awareness (Figure 1). These robots detect hostiles and injured individuals using bespoke LLMs for visual question answering (VQA), demonstrating their effectiveness in real-world deployments.

The rest of this paper is organized as follows: in the next section we review related work, followed by sections on methodology, results, and conclusions and future directions.

Related Work

Recent advancements in autonomous robotics and edge computing have highlighted the potential of Large Language Models (LLMs) to enhance decision-making and contextual

[*]These authors contributed equally.

Figure 1: Experiments with a SPOT robot.

understanding. Traditional robotic systems primarily rely on centralized processing units or cloud-based infrastructures to interpret sensor data and generate actions. However, these methods face significant challenges in terms of latency, bandwidth constraints, and dependency on stable network connectivity, particularly in dynamic and resource-constrained environments.

Edge Computing in Robotics

The deployment of edge computing solutions on robotic platforms has gained attention for its ability to process data locally, reducing latency and enhancing real-time responsiveness. Studies such as those by (Chowdhury et al. 2023), (Khan et al. 2023), and (Lee et al. 2021) have demonstrated the integration of edge-based processing with sensors like mmWave radar to improve situational awareness in adverse conditions. However, these approaches typically rely on traditional machine learning algorithms and do not leverage the contextual reasoning capabilities of LLMs. Our work addresses this gap by introducing Edge LLMs as a core processing unit for real-time decision-making in multi-modal environments.

Large Language Models for Multi-Modal Contexts

LLMs have revolutionized natural language processing by providing powerful pre-trained models capable of understanding and generating human-like text. Recent works, such as (Lewis et al. 2020) on Retrieval-Augmented Generation (RAG) and (Liu et al. 2023) on Prompt Engineering, have extended LLM applications to multi-modal tasks. These models have been effectively employed for tasks such as visual-language understanding, question answering, and contextual scene generation. However, the integration of LLMs into edge environments remains an emerging research area. Unlike traditional LLM deployments, Edge LLMs must operate under strict computational and energy constraints while maintaining high levels of inference accuracy.

Multi-Modal Sensors in Zero-Visibility Navigation

Multi-modal sensors, including mmWave radar, thermal cameras, and LiDAR, have proven critical for navigation in zero-visibility environments. mmWave radar, in particular, has been extensively studied for its robustness in adverse weather conditions and occluded environments (Lee et al. 2021; Lewis et al. 2020). (Yadav et al. 2022) demonstrated the use of mmWave radar for real-time activity recognition, while (Zhang, Chen, and Li 2024). (Liu et al. 2023) highlighted its role in collaborative multi-robot systems.

Millimeter-wave (mm-wave) radar technology has gained traction due to its ability to operate effectively in conditions where optical sensors fail, such as through snow, rain, and fog. Research by (Chowdhury et al. 2023) demonstrated moving target detection using millimeter-wave frequency-modulated continuous-wave (mmWave FMCW) radars. The study integrates classical digital signal processing (DSP) techniques, including wavelet transform, FIR filtering, and peak detection, to improve detection accuracy in dynamic outdoor environments. (Yadav et al. 2022) described mmWave radar technology for classifying human activities in real-time using edge computing for health monitoring and smart environments. Other related work includes (Lee et al. 2021) for target detection in maritime environments, (Goswami et al. 2019) for real-time multi-gesture recognition. (Lin et al. 2023) described radar-based target detection and beamforming in IoT networks respectively. However, the current work is the first in navigation in robot navigation in zero visibility conditions.

In this context, our work leverages mmWave radar as a supporting modality for Edge LLMs, enabling robust data fusion and contextual interpretation.

Human-Robot Collaboration and Decision-Making

The integration of LLMs into robotic platforms also enhances human-robot collaboration by enabling more natural and context-aware interactions. Works by (Harlow, Smith, and Thompson 2024; Radford et al. 2021; Torres, Brown, and Chen 2024) have explored the use of pre-trained models to improve interpretability and adaptability in robotic systems. Building on these efforts, we introduce an Edge LLM framework that processes real-time inputs from sensors and generates actionable insights for autonomous navigation and decision-making in zero-visibility scenarios. The integration of advanced learning paradigms is crucial for enhancing the adaptability and performance of robotic systems in dynamic environments. Table 1 offers a comprehensive comparison of Retrieval-Augmented Generation (RAG), Few-Shot Learning, Zero-Shot Learning, and Prompt Engineering, each with distinct advantages and implementation complexities:

- Retrieval-Augmented Generation (RAG): Combines retrieval and generation to enhance contextual understanding, particularly useful for knowledge-intensive tasks such as detailed question answering (Lewis et al (2020), Kelvin et al (2020), Karpukhin (2020), and Saleh (2020)).

- Few-Shot Learning: Learns from a few examples, enabling rapid adaptation to new tasks with limited data.
- Zero-Shot Learning: Generalizes from pre-trained knowledge to perform tasks without task-specific data, facilitating immediate deployment.
- Prompt Engineering: Uses structured prompts to guide the output of pre-trained models, offering flexibility in text generation tasks.

This paper bridges the gap between Edge AI and autonomous robotics by combining the strengths of LLMs and multi-modal sensing. By focusing on Edge LLMs for real-time contextual understanding, our work contributes to advancing scalable, autonomous robotic platforms capable of operating effectively in the most challenging conditions.

Applications in Surveillance, Search and Rescue

The practical applications of these technologies are vast, ranging from military operations to search and rescue missions. In military settings, the ability to navigate and identify targets in low visibility conditions is critical for operational success. Research by (Harlow, Smith, and Thompson 2024; Zhang, Chen, and Li 2024; Li, Wang, and Zhao 2024) on using mm-wave radar and multi-robot collaboration in tactical environments demonstrated significant improvements in mission outcomes. Similarly, in search and rescue scenarios, the prompt identification of injured individuals and efficient navigation through debris-laden environments can save lives, as evidenced by numerous field trials and simulations.

Table 1 compares four approaches—RAG (Retrieval-Augmented Generation), Few-Shot Learning, Zero-Shot Learning, and Prompt Engineering—based on their core ideas, knowledge sources, use cases, contextual understanding, and implementation complexity. RAG combines retrieval and generation, leveraging external knowledge bases for knowledge-intensive tasks, such as answering medical questions using literature, but requires higher implementation complexity due to retrieval integration (Lewis et al. 2020). Few-Shot Learning adapts to new tasks using a small number of annotated examples, balancing moderate implementation complexity with flexibility for tasks like language translation with limited data (Brown et al. 2020). Zero-Shot Learning generalizes from pre-trained knowledge to perform tasks without task-specific data, offering immediate adaptability and scalability for general knowledge tasks (Radford et al. 2021) Prompt Engineering relies on well-crafted structured prompts to guide pre-trained models for diverse conversational and creative tasks, requiring minimal implementation complexity but limited by the model's pre-trained knowledge (Liu et al. 2023). While RAG excels in knowledge-intensive tasks, Few-Shot and Zero-Shot Learning provide adaptability for varied scenarios, and Prompt Engineering supports flexible text generation.

Methodology

Our methodology includes data preparation, model setup, context processing, response generation, and evaluation. Thermal images and pre-trained BERT (Devlin et al. 2019) models are used for visual question answering tasks. Cosine similarity metrics evaluate response accuracy, distinguishing between relevant and irrelevant answers.

Data Preparation

The first step involves preparing both image and textual data. For this research, we use thermal images that depict various battlefield scenarios, thus providing a clear context for each image. For textual data, we curate ground truth responses for each image context, which serve as benchmarks for evaluating the model's performance. Additionally, we create various types of responses, including correct, incorrect, irrelevant, and gibberish responses, to comprehensively test the model's ability to distinguish between relevant and irrelevant answers.

Model Setup

In Algorithm 1 we present RAG++, a novel algorithm designed to evaluate the similarity between response sentences and a predefined ground truth using BERT embeddings. The method begins by tokenizing both the ground truth and response sentences, converting them into embeddings via a pre-trained BERT model. These embeddings are then averaged using mean pooling to produce a single vector representation for each sentence. The algorithm proceeds by computing the cosine similarity between each response vector and the ground truth vector, offering a quantitative measure of semantic similarity. This approach is particularly effective for tasks requiring a detailed comparison of textual responses, providing a robust framework for evaluating the relevance and accuracy of generated text in natural language processing (NLP) applications.

The core of our methodology is the pre-trained BERT model, specifically the 'bert-base-uncased' model from Hugging Face's Transformers library. This model is selected due to its strong performance in natural language understanding tasks. Alongside the model, we use the corresponding 'BertTokenizer' to tokenize input texts, ensuring consistency with the model's training setup. This setup allows us to leverage the powerful contextual understanding capabilities of BERT for our VQA tasks.

Context Processing

In context processing, each textual context associated with an image is tokenized using the 'BertTokenizer'. This process involves converting text into token IDs and creating attention masks to identify relevant tokens, ensuring that the input is properly formatted for the BERT model. Once tokenized, the inputs are passed through the pre-trained BERT model to obtain hidden states. These hidden states serve as the foundational representation of the text, capturing the nuanced meaning and context necessary for generating accurate responses.

Response Generation

Based on the model's outputs, we generate responses for each context. This involves interpreting the hidden states produced by the BERT model to formulate coherent and

Aspect	RAG	Few-Shot	Zero-Shot	Prompt Eng.
Knowledge Source	External	Few Examples	Pre-trained Model	Pre-trained Model
Adaptability	High	Moderate	High	Moderate
Complexity	High	Moderate	Low	Low
Best Use Case	Knowledge Retrieval	Task Adaptation	Generalization	Structured Prompts

Table 1: Comparison of Learning Approaches for Edge LLMs

(a) Soldiers with weapons in a tunnel

(b) Armed civilian group clustering

(c) Injured individuals on the ground

Figure 2: The three contexts used in the experiments.

contextually appropriate answers. Various types of responses are generated, including correct, incorrect, irrelevant, and gibberish, to evaluate the model's ability to discern and produce relevant information in different scenarios.

Algorithm 1, titled **RAG++ Using BERT Embeddings and Cosine Similarity**, presents a systematic approach for evaluating the semantic similarity between a predefined ground truth sentence and multiple response sentences. This method leverages pre-trained BERT embeddings and cosine similarity as a metric to quantify the closeness of each response sentence to the ground truth. The algorithm is designed for tasks requiring fine-grained semantic analysis and comparison of textual inputs, such as response evaluation in natural language processing systems.

The process begins by initializing the necessary components, including a pre-trained BERT model, its tokenizer, and essential libraries such as 'sklearn' for cosine similarity computation and 'numpy' for numerical processing. The input consists of a selected ground truth sentence and a set of six response sentences. These textual inputs are tokenized using the BERT tokenizer with configurations such as truncation, padding, and a maximum token length of 512 to ensure compatibility with the model's architecture.

Following tokenization, the algorithm disables gradients to optimize computation and passes the tokenized inputs through the BERT model. The embeddings are derived from the last hidden state of the model, where mean pooling is applied to obtain fixed-length vector representations for each sentence. These embeddings are subsequently converted into NumPy arrays to facilitate efficient numerical operations.

The final step involves calculating cosine similarity scores between the embedding of the ground truth sentence and the embeddings of the response sentences. This similarity metric provides a normalized measure of semantic alignment, ranging between -1 and 1, with higher values indicating closer alignment. The computed similarity scores are printed for each response sentence, enabling a comprehensive evaluation of their semantic correspondence with the ground truth.

This algorithm provides a robust and scalable method for assessing textual similarity in various applications, such as response evaluation in question-answering systems or dialogue systems. By integrating pre-trained language models and efficient similarity computation, it demonstrates the potential for high-precision semantic analysis in real-world scenarios.

Algorithm 1: RAG++ Using BERT Embeddings and Cosine Similarity

Input: Ground truth sentence GT, Response sentences $\{T_1, T_2, ..., T_6\}$
Output: Cosine similarity scores between each response and GT
Initialize BERT model and tokenizer
Define ground truth and response sentences
for each response T_i **do**
    Tokenize T_i and GT
    Compute embedding E_i using BERT
end for
Compute ground truth embedding E_{GT}
for each embedding E_i **do**
    Compute cosine similarity: $S_i = \text{cosine\_similarity}(E_i, E_{GT})$
    Print S_i
end for

Algorithm 2: Compare Different Embedding Methods

Require: Ground Truth Sentence GT, Text Sentences T, Embedding Method $\mathcal{M}$
Ensure: Cosine Similarity Scores for each sentence in T
Initialize pre-trained tokenizer and model
Tokenize GT and each $T_i \in T$
Pass tokenized inputs through the model to obtain hidden states
Compute embeddings using method $\mathcal{M}$:
for each T_i **do**
    Compute $E_i = \mathcal{M}(T_i)$
end for
Compute ground truth embedding: $E_{GT} = \mathcal{M}(GT)$
for each E_i **do**
    Compute cosine similarity: $S_i = \text{cosine\_similarity}(E_i, E_{GT})$
    Print S_i
end for

Evaluation

Algorithm 2, titled **Compare Different Embedding Methods**, extends the approach of Algorithm 1 by introducing flexibility in generating embeddings for textual inputs using various strategies. It allows for experimentation with multiple embedding techniques, including mean pooling, max pooling, and attention-based methods, to compute cosine similarity between a ground truth sentence and a list of response sentences. This method is particularly valuable in contexts where different embedding strategies may yield improved performance based on the characteristics of the dataset or task.

The algorithm begins by initializing the required components, including a pre-trained tokenizer and language model, as well as specifying the embedding method to be used. The input includes a predefined ground truth sentence and a list of response sentences. Tokenization is performed for both the ground truth and response sentences using the pre-trained tokenizer, ensuring compatibility with the model's input requirements, such as truncation, padding, and a maximum token length of 512.

Once tokenized, the textual inputs are passed through the language model to obtain the hidden states. Depending on the chosen embedding method, different strategies are applied to generate the sentence embeddings. For example, mean pooling and max pooling aggregate information across token embeddings in different ways, while attention-based methods extract embeddings from the last layer or the last four layers of the model. This modular approach allows for tailored embedding generation suited to specific use cases.

After obtaining embeddings for the response sentences and the ground truth, cosine similarity is computed to measure the semantic alignment between them. The similarity scores are then printed, providing an evaluation of how closely each response sentence corresponds to the ground truth in the embedding space.

Algorithm 2 highlights the adaptability of embedding methods in natural language processing tasks. By offering multiple strategies for embedding generation, it enables researchers and practitioners to explore and compare the effectiveness of different approaches for a given application. This flexibility ensures that the algorithm can be fine-tuned to meet the requirements of diverse textual analysis tasks, ranging from semantic similarity measurement to content classification and retrieval.

The generated responses are evaluated against the ground truth using cosine similarity. This metric measures the closeness of the generated response to the expected response in the vector space, providing a quantitative assessment of accuracy and relevance. Performance metrics are established by categorizing the responses into correct, incorrect, irrelevant, and gibberish. The cosine similarity scores for each category are analyzed to determine the model's effectiveness in generating context-appropriate responses. This thorough evaluation process ensures that the VQA system is rigorously tested and provides valuable insights into its performance and areas for improvement.

By following this methodology, we systematically evaluate the VQA system's performance across different contexts, ensuring that previous contexts do not influence the results of new contexts. This approach highlights the model's ability to understand and respond appropriately to diverse scenarios, providing valuable insights into its effectiveness and areas for improvement.

Results

Experiments demonstrate the effectiveness of the proposed system in various contexts, including identifying threats and injured individuals under zero-visibility conditions. Results show significant improvements in response accuracy and contextual understanding.

Experimental Setting

A multimodal LLM based on LLaVA 1.5 was deployed on a Jetson Orin NX, a compact edge device featuring up to 6 ARM v8.2 CPU cores, an NVIDIA Ampere GPU with 1024

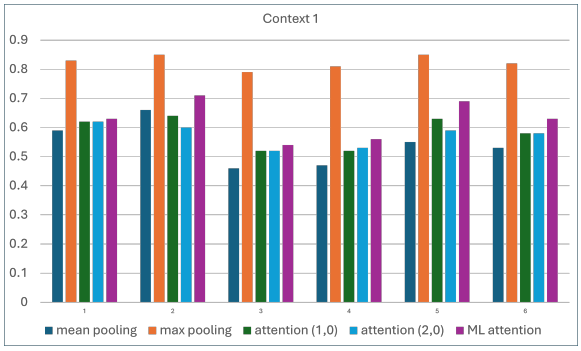
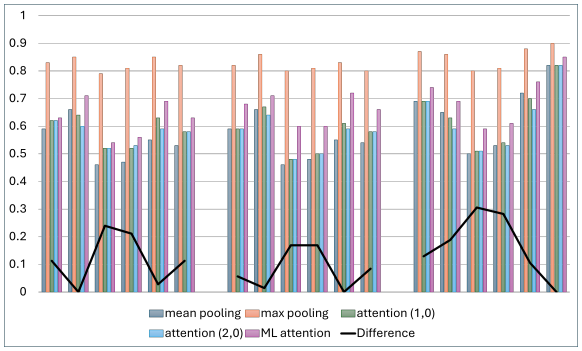
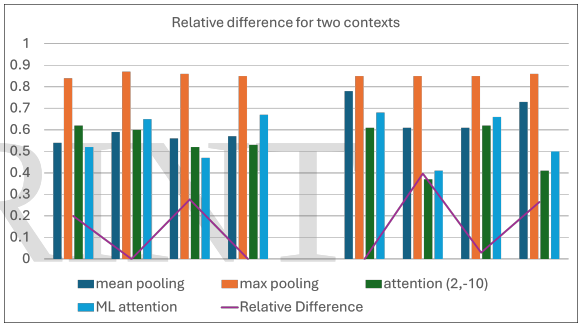


Figure 3: Comparison of results.



(a)



(b)

Figure 4: Comparative analysis of various embedding methods.

CUDA cores, 16GB LPDDR5 memory, and 100 TOPS of AI performance. LLaVA 1.5 was fine-tuned on 1,500 military scenario images collected during testing. For baseline comparisons and to avoid constraints from limited CPU or memory resources, other LLMs were deployed on a dedicated edge server equipped with an RTX 4090 GPU (24GB), a high-core-count CPU, and 64GB RAM. This dual setup enabled a comparative evaluation of fine-tuned LLaVA 1.5 under realistic edge conditions versus a high-capacity server.

The experiments assessed the RAG++ framework's efficacy for zero-visibility navigation and contextual scene understanding in battlefield scenarios. Thermal images simulated environments with visual obstructions like smoke and low light, covering contexts such as armed soldiers in tunnels, injured individuals, and clustering combat groups. Each context was paired with a ground truth description, and system-generated responses were evaluated using cosine similarity metrics.

Pre-trained language models were used for embedding generation, while visual-language models (VLMs) generated contextual descriptions. Responses were categorized as correct, incorrect, irrelevant, or partially true to test the system's ability to differentiate relevant information. Performance was analyzed across three scenarios, demonstrating the robustness and adaptability of the framework.

Analysis

The results highlight the system's capability to generate contextually relevant responses across diverse scenarios. For **Context 1** - (soldiers with weapons in a tunnel) shown in Figure 2 (a), the system correctly identified enemy soldiers with weapons and accurately recommended a counteroffensive response. However, irrelevant and partially correct responses such as "This is a scene from a dance party" or "People with weapons indicating civilian unrest" were also observed, indicating areas for improvement in response generation.

In **context 2** - (grouping of armed groups), the system correctly identified the presence of an armed group preparing for coordinated action shown in Figure 2 (b). Partially correct responses such as "Enemy soldiers with weapons in a battlefield" were generated, which, while accurate in some aspects, lacked specificity to the context. Irrelevant responses like "No immediate threat detected" were penalized, emphasizing the need for enhanced contextual alignment.

For **Context 3**- (injured individuals on the ground), shown in Figure 2 (c), the system successfully identified the need for medical assistance and generated the correct response: "Injured individuals needing medical assistance. " However, irrelevant responses such as "This is a scene from a soccer match" and incorrect interpretations such as "Enemy soldiers with weapons on a battlefield" highlighted challenges in adapting to highly specific contexts.

Quantitative Evaluation

The system's performance was quantitatively evaluated by calculating the cosine similarity scores between response embeddings and the ground truth embeddings. Correct responses consistently achieved the highest similarity scores, while irrelevant and incorrect responses scored significantly lower. Across all three contexts, the RAG++ framework demonstrated higher accuracy compared to baseline methods such as prompt engineering and few-shot learning. The following responses were used for the quantitative evaluation of the RAG++ algorithm.

- **R1:** Enemy soldiers with weapons in a battlefield.
- **R2:** This is a scene from a soccer match.
- **R3:** This is a scene from a dance party.
- **R4:** This is a scene from a soccer match.
- **R5:** People with weapons indicating civilian unrest.

	Jetson Orin AGX (64 GB)	Fog Server (4090 RTX)	Jetson Orin Nano (8GB)
Idle Power Performance	9.6 W	320 W	7.6 W
Peak Performance	32.1 W	560 W	23 W
Model Output	19.25 token/sec	124.63 token/sec	4.6 token/sec
Latency - Audio (WiFi)	1.2 sec	1.2 sec	2 sec
Power Source	Robot Battery	External Power	Robot Battery
RAM Usage with Torch	9.1 GB	12 GB	6.8 GB

Table 2: Performance Monitoring of Edge and Fog Servers

- **R6:** Injured individuals needing medical assistance.

As explained in Algorithms 1 and 2, we experiments with various embedding methods including *mean pooling*, *max pooling*, and *attention* model with the output feature maps of the last layer, and attention model with the last four layers (ML_attention). In addition, we experimented with two attention models with various attention weights to the regions of interest. The results are shown in Figure 4. The six responses (R1-R6) were evaluated against the ground truth using cosine similarities for each embedding method discussed above. Max pooling consistently overestimated the similarities for all responses, showing very little discriminatory power. The multi-layer attention model performed the best in terms of its discriminatory power between correct and incorrect responses.

Figure 4a extended the comparison of the five embedding methods in terms of the relative and absolute differences between the ground truth and responses given. The relative difference is the difference in the confidence scores of each response from the correct response. The absolute difference is the same as the relative difference except for the fact that he correct response is assigned a confidence score of 1. Figure 4b shows results from a second set of experiments where longer responses were penalized for verbosity by assigning negative weights to irrelevant words. RAG++ showed the minimum verbosity followed by few-shot learning. Both pre-trained model and prompt engineering produced longer responses. In critical situations such as search and rescue or military operations verbose responses are problematic and hence we tested the verbosity of responses. The line shown in both Figures 4a and 4 plot the differences and show that the multi-layer attention model has the highest discriminatory power.

Real-Time Tracking and Monitoring

Each robot's location is tracked using a Garmin 18x module mounted on the Jetson Orin Nano, which serves as the robot's payload and is powered by its internal battery. The location data is continuously streamed to the dashboard via a message queue and simultaneously stored in MongoDB to enable route history tracking for efficient path optimization. MongoDB is hosted on a Fog server and communicates with the Jetson using a client-server architecture.

The performance of the system is shown in Table 2. We ran LLaVA 7B on both the Jetson Orin AGX and the Fog server, while VILA 2.7B was deployed on the Jetson Orin Nano. The Orin AGX was preferred for its efficient token generation and optimized power consumption compared to the Jetson Nano. Additionally, we evaluated an object detection model on both the AGX and Nano, observing that the Nano struggled to sustain real-time performance.

Conclusions and Future Work

In this paper, we presented a novel framework integrating Edge Large Language Models (LLMs) into robotic platforms for real-time decision-making and contextual understanding in zero-visibility environments. By leveraging the processing capabilities of Edge LLMs, our system demonstrated significant improvements in interpreting multi-modal inputs, such as thermal images and mmWave radar data, and generating semantically accurate responses. The experiments showcased the framework's adaptability and robustness in dynamic scenarios, including tactical operations and disaster response, where traditional sensor modalities often fail. The combination of Edge LLMs and multi-modal sensing enabled autonomous robotic platforms to prioritize tasks, detect threats, and provide actionable insights in real-time, reducing dependency on high-bandwidth communication links.

The results validate the feasibility of deploying LLMs on resource-constrained edge devices, marking a step forward in advancing autonomous systems. Furthermore, the use of cosine similarity metrics for evaluating the semantic alignment of generated responses highlighted the framework's ability to balance computational efficiency with accuracy, even under challenging environmental conditions.

Despite these achievements, several areas remain for future exploration. One promising direction is the application of Edge LLMs to **swarm robotics**, where multiple robots collaborate to achieve complex objectives. The integration of LLMs into swarm systems could enhance communication, coordination, and decision-making by providing a unified language-based interface for sharing contextual insights. This would enable more intelligent and adaptable behaviors, such as dynamic task allocation, environmental mapping, and coordinated navigation in unstructured terrains.

Additionally, further optimization of Edge LLMs is required to meet the energy and computational constraints of swarm robotics. Techniques such as model pruning, quan-

tization, and distillation could reduce the resource footprint while maintaining performance. Incorporating reinforcement learning and adaptive fine-tuning mechanisms could also enable Edge LLMs to evolve in real-time, improving their responsiveness to changing conditions.

Future work will also focus on expanding the framework to support a broader range of sensor modalities, such as LiDAR, acoustic sensors, and hyperspectral imaging, for richer multi-modal fusion. Enhanced evaluation metrics beyond cosine similarity, such as task-specific performance measures, will be explored to provide a more comprehensive assessment of the system's capabilities.

We demonstrate that the integration of Edge LLMs into robotic platforms represents a transformative step for autonomous systems. Extending this innovation to swarm robotics holds immense potential for creating scalable, intelligent, and resilient robotic networks capable of operating effectively in the most challenging environments.

Acknowledgments

This work is supported by U.S. Army Grant No: W911NF2120076.

References

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33: 1877–1901.

Chowdhury, D.; Melige, N. V.; Pal, B.; and Gangopadhyay, A. 2023. Enhancing outdoor moving target detection: Integrating classical DSP with mmWave FMCW radars in dynamic environments. *Electronics*, 12(24): 5030.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186. Association for Computational Linguistics.

Goswami, P.; Rao, S.; Bharadwaj, S.; and Nguyen, A. 2019. Real-time multi-gesture recognition using 77 GHz FMCW MIMO single chip radar. In *Proceedings of the 2019 IEEE International Conference on Consumer Electronics (ICCE)*, 1–4. Las Vegas, NV, USA.

Harlow, J.; Smith, E.; and Thompson, L. 2024. A New Wave in Robotics: Survey on Recent mmWave Radar Applications in Robotics. *arXiv preprint*.

Khan, M.; Ahmed, N.; Padela, J.; Raza, M.; Gangopadhyay, A.; Wang, J.; Foulds, J.; Busart, C.; and Erbacher, R. 2023. Flood-ResNet50: Optimized Deep Learning Model for Efficient Flood Detection on Edge Device. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, 512–519. IEEE.

Lee, M. J.; Kim, J. E.; Ryu, B. H.; and Kim, K. T. 2021. Robust maritime target detector in short dwell time. *Remote Sensing*, 13: 1319.

Lewis, P.; Perez, E.; Piktus, A.; Karpukhin, V.; Goyal, N.; Kulshreshtha, A.; Min, S.; Yih, W.-t.; Yuan, X.; Constant, N.; et al. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, 9459–9474.

Li, Z.; Wang, Y.; and Zhao, M. 2024. A Robust Multiobject Tracking Method Based on 4-D Millimeter-Wave Radar and Monocular Vision Fusion. *IEEE Transactions on Intelligent Vehicles*.

Lin, Z.; Niu, H.; An, K.; Hu, Y.; Li, D.; Wang, J.; and Al-Dhahir, N. 2023. Pain without gain: Destructive beamforming from a malicious RIS perspective in IoT networks. *IEEE Internet of Things Journal*. Advance online publication.

Liu, P.; You, Z.; Zhang, X.; and Chen, J. 2023. Prompt Engineering: Optimizing Pre-trained Language Models for Diverse Applications. *Journal of Machine Learning Research*, 24(7): 1–24.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *Proceedings of ICML*, 8748–8763.

Torres, J.; Brown, D.; and Chen, Y. 2024. Gesture and Voice Command-Based Control for Edge AI Robots. *ACM Transactions on Human-Robot Interaction*, 15(1): 1–20.

Yadav, S. S.; Agarwal, R.; Bharath, K.; Rao, S.; and Thakur, C. S. 2022. TinyRadar: MmWave radar based human activity classification for edge computing. In *Proceedings of the 2022 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2414–2417. Austin, TX, USA.

Zhang, W.; Chen, H.; and Li, X. 2024. A Robust Robot Perception Framework for Complex Environments Using Multiple mmWave Radars. *IEEE Transactions on Robotics*.