INTERNATIONAL JOURNAL OF APPLIED METHODS IN ELECTRONICS AND COMPUTERS

www.ijamec.org

International Open Access

Volume 14
Issue 01
March, 2026

Research Article

Domain-Specific Fine-Tuning of LLMs for Elevator Emergency Response Systems

Busra Onal [a], ***Muhammet Fatih Aslan*** [b,*], ***Akif Durdu*** [c]

[a] *Butkon Elevator Industry Trade Inc., Research and Development Center, Konya, Türkiye*
[b] *Department of Artificial Intelligence and Machine Learning, Faculty of Computer and Information Sciences, Konya Technical University, Konya, Türkiye*
[c] *Department of Electrical-Electronic Engineering, Konya Technical University, Konya, Türkiye*

ARTICLE INFO

Article history:
Received 3 February 2026
Accepted 25 March 2026

Keywords:
Edge AI,
Elevator Safety,
Fine-Tuning,
Large Language Models,
QLoRA.

ABSTRACT

Large Language Models (LLMs) have demonstrated remarkable capabilities in natural language understanding and generation. However, their application in safety-critical domains such as elevator emergency response requires domain-specific knowledge and reliable performance, particularly when deployed as offline edge AI systems with constrained computational resources. This study presents fine-tuning of the Gemma-2-9B-Instruct model using Quantized Low-Rank Adaptation (QLoRA) for Turkish elevator emergency scenarios, targeting deployment on embedded touchscreen control panels inside elevator cabins. Unlike cloud-based systems leveraging internet connectivity for retrieval, our edge deployment operates entirely offline, making fine-tuning essential for encoding domain knowledge directly into model parameters. We developed a specialized Turkish dataset containing 1,155 question-answer pairs covering diverse emergency situations. Our evaluation demonstrates significant improvements: ROUGE-1 scores increased from 0.259 to 0.317 (22.49%), BLEU improved by 218.11%, and hallucination rates reduced from 54% to 22% while achieving 49% faster inference. The resulting 4.8GB quantized model runs entirely on embedded hardware without network dependencies, providing immediate, reliable emergency guidance. These results validate parameter-efficient fine-tuning for safety-critical edge AI applications.

1. Introduction

Large Language Models (LLMs) have significantly advanced natural language processing by enabling machines to generate coherent, context-aware, and semantically meaningful text across a wide range of tasks. These capabilities are primarily driven by transformer-based architectures and large-scale pretraining on heterogeneous textual corpora, which allow models to capture linguistic structure, semantic relationships, and contextual dependencies across domains [1, 2]. Subsequent developments in instruction tuning and alignment techniques have further improved the ability of LLMs to follow user intent and produce structured responses suitable for interactive applications [3]. With the emergence of open-weight model families such as LLaMA [4] architectures and the Gemma [5] series, the deployment of LLM-based intelligent assistants has expanded beyond research prototypes into practical industrial experimentation. Nevertheless, most existing research and deployment scenarios remain focused on general-purpose conversational systems rather than domain-constrained environments where reliability, response structure, and deterministic behavior are essential.

Deploying LLMs in safety critical industrial environments introduces challenges that differ fundamentally from those encountered in conventional

* Corresponding author. E-mail address: *mfaslan@ktun.edu.tr*
DOI: 10.58190/ijamec.2026.165

conversational AI systems. In such environments, incorrect or misleading responses may create operational risks rather than minor usability problems. One of the most widely documented reliability issues in generative language models is hallucination, where models produce fluent but factually incorrect information with high confidence [6]. This phenomenon becomes particularly problematic when models are applied to specialized technical domains or low-resource language settings, where domain terminology, procedural knowledge, and regulatory information are insufficiently represented in pretraining corpora[7]. Industrial safety applications require communication that is not only linguistically correct but also procedurally accurate, concise, and actionable. Consequently, ensuring factual reliability, predictable response structure, and domain-appropriate communication becomes a central requirement when integrating LLM based assistants into safety-critical systems.

Elevator emergency response systems provide a representative example of such constrained interaction scenarios. Conventional emergency communication mechanisms (including alarm buttons, printed instructions, and intercom systems) offer static guidance and cannot dynamically adapt to user uncertainty during stressful situations. During emergency events such as entrapment, fire, earthquake, or medical incidents, occupants may experience panic, reduced situational awareness, and difficulty interpreting written instructions. An embedded AI assistant capable of delivering structured emergency guidance in natural language could improve safety outcomes by providing immediate procedural instructions and reassurance. However, implementing such functionality requires addressing deployment constraints that differ substantially from those assumed in typical cloud-based AI applications.

Unlike most contemporary LLM deployments, elevator cabin systems must operate as offline edge AI solutions. Elevators frequently function in connectivity-limited environments such as building shafts, underground parking areas, and reinforced structural cores, where wireless communication may be unreliable or unavailable. Emergency conditions may further involve communication failures, making reliance on external services unacceptable. At the same time, embedded cabin control panels operate under strict hardware limitations, typically providing only a few gigabytes of memory and low-power processors. These constraints make conventional cloud inference pipelines and retrieval-dependent architectures impractical, motivating research on parameter-efficient adaptation, quantization, and compact deployment strategies that enable fully self-contained AI assistants.

Retrieval-Augmented Generation (RAG) has emerged as a widely adopted approach for improving factual consistency in LLM-based systems by grounding responses in external knowledge sources [8]. By retrieving relevant documents during inference, RAG systems can reduce hallucination and improve factual alignment with authoritative information. However, retrieval pipelines introduce additional latency, memory overhead, and infrastructure dependencies that may conflict with real-time safety-critical interaction requirements on embedded hardware. Furthermore, fragmented user queries typical of panic situations may reduce retrieval effectiveness. These limitations motivate investigation of alternative approaches that maintain reliability while enabling deterministic and fully offline operation in resource-constrained environments.

Fine-tuning is a widely used approach for adapting pretrained large language models to downstream tasks by updating model parameters using task-specific data. Rather than training models from scratch, fine-tuning allows previously learned linguistic and semantic representations to be reused while specializing model behavior for new applications. Parameter-efficient fine-tuning methods, such as Low-Rank Adaptation (LoRA) [9], enable this adaptation with minimal computational cost by updating only a small subset of model parameters. Quantized Low-Rank Adaptation (QLoRA) [10] further extends this idea by combining LoRA with low-precision quantization, allowing large models to be fine-tuned and deployed under strict memory constraints while maintaining performance. A specialized form of this process, known as domain-specific fine-tuning, focuses on adapting language models to a particular knowledge domain by embedding domain terminology, procedural knowledge, and communication patterns directly into model parameters. Prior research in specialized domains (BioBERT [11] , LEGAL-BERT [12], etc.) demonstrates that domain-adapted models can achieve improved terminology usage, factual reliability, and task performance compared to general-purpose LLMs. These findings suggest that parameter-efficient domain-specific fine-tuning is particularly well suited for safety-critical edge AI applications, where computational efficiency and domain accuracy must be balanced.

In this work, we investigate QLoRA-based fine-tuning of the Gemma-2-9B-Instruct model for Turkish elevator emergency response scenarios, targeting deployment on embedded touchscreen control panels inside elevator cabins. A domain-specific dataset consisting of 1,155 Turkish question–answer pairs was constructed using elevator safety manuals, incident reports, and expert validation. The adapted model is deployed using 4-bit quantization, enabling fully offline execution on resource-constrained hardware[10]. Experimental results demonstrate improvements in response quality, reduction in hallucination rates, and faster inference compared to the base model, indicating that parameter-efficient domain

adaptation can enable reliable LLM deployment in safety-critical industrial environments. The primary contributions of this work are:

Development of a domain-specific Turkish elevator emergency response dialogue dataset covering diverse real-world emergency scenarios, validated by certified safety professionals against international standards.

Parameter-efficient fine-tuning of a 9B-parameter instruction-tuned LLM using QLoRA, enabling effective domain adaptation while updating less than 1% of model parameters.

Offline deployment of a fully quantized LLM on embedded hardware, eliminating network dependency for safety-critical real-time emergency guidance.

A comprehensive evaluation combining generation-quality, reliability, and efficiency metrics, including hallucination rate assessment tailored for safety-critical applications.

2. Related Work

LLMs often require adaptation to specialized domains in order to achieve reliable performance. Although general-purpose models are trained on large and diverse corpora, they may struggle with domain-specific terminology due to domain shift in specialized texts. To address this problem, prior research has adopted domain-specific pre-training and fine-tuning strategies. For example, BioBERT[11], developed for biomedical text mining, was trained on biomedical corpora and demonstrated improved performance over the general BERT model in named entity recognition (NER) and relation extraction (RE) tasks. Similarly, SciBERT[13] achieved performance improvements for scientific text processing by introducing a domain-specific vocabulary (SCIVOCAB). ClinicalBERT[14], trained on clinical notes, and LEGAL-BERT[12], designed for legal documents, further demonstrate the importance of domain adaptation in specialized language modeling. In the financial domain, FinBERT[15] has shown improved performance in financial sentiment analysis compared to general-purpose language models.

To adapt large language models efficiently, Parameter-Efficient Fine-Tuning (PEFT) methods have gained popularity as an alternative to full-model fine-tuning. LoRA reduces training cost by freezing pretrained model weights and injecting trainable low-rank matrices into transformer layers, significantly reducing the number of trainable parameters. A more recent approach, QLoRA, further improves memory efficiency by combining LoRA with low-precision quantization. QLoRA introduces innovations such as the 4-bit NormalFloat data type and double quantization, enabling fine-tuning of models with tens of billions of parameters on limited hardware without significant performance degradation.

Research on the use of LLMs in elevator systems remains limited, although recent studies have begun to explore maintenance and technical-support applications. A recent study focusing on elevator maintenance strategies demonstrated that large language models can be adapted using domain-specific maintenance data and that LoRA-based fine-tuning can improve diagnostic accuracy [16]. The study emphasizes that reliable information generation is critical in maintenance workflows and that hallucination remains an important safety concern. Another study investigated the use of GPT-4-based LLMs as evaluation tools for task-oriented industrial chatbots [17]. The results indicate that LLM-based evaluation aligns well with human judgment for low-complexity tasks but becomes less reliable for higher cognitive complexity scenarios. These findings highlight the importance of reliability assessment when deploying LLM-based systems in industrial environments.

Existing literature indicates that LLM applications in elevator systems primarily focus on maintenance optimization and product-selection assistance. However, offline emergency-guidance assistants operating inside elevator cabins on embedded hardware remain largely unexplored. This study aims to address this gap by developing an elevator emergency assistant capable of running on resource constrained hardware using the memory efficient QLoRA fine-tuning approach, enabling reliable emergency guidance without requiring internet connectivity.

3. Methodology

This study aims to develop a domain-adapted large language model capable of providing reliable emergency guidance within elevator cabins under offline and hardware-constrained conditions. Unlike cloud-based conversational systems, the proposed approach focuses on embedding procedural knowledge directly into model parameters through parameter-efficient fine-tuning. The methodological framework is designed to ensure three primary objectives: (i) domain specialization for Turkish elevator emergency scenarios, (ii) memory-efficient adaptation compatible with embedded deployment, and (iii) reduction of hallucination risks in safety-critical responses.

The overall methodology consists of four sequential stages. First, a structured domain-specific dataset was constructed based on realistic elevator emergency scenarios. Second, a pretrained instruction-tuned large language model (Gemma2-9B-Instruct) was selected as the base model. Third, parameter-efficient fine-tuning was applied using the QLoRA method to adapt the model to the emergency domain. Finally, the trained model was evaluated and prepared for deployment in a quantized format suitable for embedded systems. The following

subsections describe each stage in detail.

3.1. Dataset Construction

A domain-specific Turkish dataset consisting of 1,155 question–answer pairs was developed to represent realistic elevator emergency scenarios. Data collection involved a systematic review of safety manuals from major elevator manufacturers (Otis, Schindler, KONE, ThyssenKrupp), consultation with certified elevator safety professionals, and analysis of documented incident reports. The objective was to construct a dataset that reflects procedurally accurate, safety-compliant, and context-aware emergency guidance suitable for deployment in elevator cabin environments.

The dataset covers seven primary categories: (1) mechanical failures, (2) entrapment situations, (3) fire and smoke incidents, (4) earthquake-related scenarios, (5) medical emergencies, (6) safety concerns such as air supply and structural integrity, and (7) communication failures. Each response was written in a concise, directive format prioritizing clarity, actionability, and reassurance, critical characteristics for crisis communication. All samples follow a structured instruction, response format compatible with instruction-tuned transformer models. A representative example is shown in Figure 1.

Figure 1. Sample instruction-response structure used for fine-tuning (Turkish original (left) and English translation (right).)

To ensure dataset reliability and safety compliance, a multi-stage quality assurance process was applied. This included expert review by certified elevator safety professionals, linguistic quality control by native Turkish speakers, automated consistency checks through validation scripts, and safety verification against international standards, including EN 81 and ASME A17.1. These procedures were implemented to minimize ambiguity, eliminate procedural inconsistencies, and reduce the risk of unsafe or misleading guidance.

The dataset was divided into training and testing subsets using stratified sampling to preserve category balance. An 85%/15% split was applied, resulting in 1,039 training samples and 116 test samples. This stratified approach ensured that all emergency categories were proportionally represented in both subsets, supporting fair evaluation of model performance.

3.2. Model: Gemma-2-9B-Instruct

As the base model for domain adaptation, Gemma-2-9B-Instruct [5] was selected due to its balance between performance, architectural efficiency, and compatibility with parameter-efficient fine-tuning methods. Gemma-2 is a transformer-based large language model trained using large scale corpora and subsequently instruction-tuned to improve task-following and conversational alignment. The 9-billion-parameter variant offers sufficient representational capacity for domain adaptation while remaining feasible for quantized deployment on resource-constrained hardware.

The model follows a decoder-only transformer architecture with multi-head self-attention layers and feed-forward networks optimized for autoregressive text generation. Its instruction-tuned configuration allows it to interpret structured prompt formats, including role-based conversational inputs, making it suitable for the emergency guidance dialogue structure introduced in Section 3.1. Since the dataset was constructed in a user–assistant message format, alignment with the model's original instruction-tuning objective was preserved during adaptation.

One of the key reasons for selecting Gemma-2-9B-Instruct instead of larger-scale models (e.g., 30B+ parameters) was deployment feasibility. The target system is an embedded elevator cabin control panel with strict memory and computational constraints. Larger models would require either multi-GPU infrastructure or aggressive compression strategies that could negatively impact reliability. Conversely, smaller models (e.g., <3B parameters) may lack sufficient representational capacity to internalize procedural safety instructions effectively. Therefore, the 9B configuration represents a compromise between domain adaptation capability and edge deployment practicality.

Before fine-tuning, the pretrained weights were loaded in 4-bit precision format to enable memory-efficient training using QLoRA. The base model parameters remained frozen during adaptation, ensuring that domain knowledge was incorporated exclusively through lightweight adapter layers. This design choice reduces catastrophic forgetting and preserves the general linguistic competence of the pretrained model while injecting elevator-specific emergency knowledge.

3.3. QLoRA Configuration

To enable memory-efficient domain adaptation of the Gemma-2-9B-Instruct model, QLoRA[10] was employed. The complete configuration and resource footprint are summarized in Table 1. As shown in Table 1, the pretrained model weights were loaded in 4-bit NormalFloat (NF4) precision with double quantization

enabled, while training computations were performed in bfloat16 format to ensure numerical stability. Instead of updating all model parameters, LoRA adapters were injected into the transformer attention projection layers, specifically the q_proj, k_proj, v_proj, and o_proj modules.

Table 1. QLoRA Configuration Parameters

Category	Parameter	Value
Base Model	Model Name	Gemma-2-9B-Instruct
Quantization	Weight Precision	4-bit NF4
	Double Quantization	Enabled
	Compute Data Type	bfloat16
LoRA Configuration	Rank (r)	32
	Alpha (α)	64
	Dropout	0.05
Target Modules	Injected Layers	q_proj, k_proj, v_proj, o_proj
Training Output	Trainable Parameters	37M (~0.4%)
Memory Footprint	Adapter Size	148 MB
Deployment Size	Final Quantized Model	4.8 GB

The LoRA configuration was defined with a rank (r) of 32, a scaling factor (α) of 64, and a dropout rate of 0.05. Under this setup, only 37 million parameters were updated during training, corresponding to approximately 0.4% of the total model parameters. The resulting adapter size was 148 MB, while the fully quantized deployed model occupied 4.8 GB of memory. This configuration significantly reduces memory requirements compared to full fine-tuning while preserving sufficient adaptation capacity for domain-specific learning. As indicated in Table 1, the final model size and adapter footprint make the system suitable for deployment in resource-constrained embedded environments such as elevator cabin control units.

3.4. Training

The model was fine-tuned using the domain-specific dataset described in Section 3.1 and the QLoRA configuration presented in Table 1. The complete training setup and outcomes are summarized in Table 2. During preprocessing, the Gemma-2 chat template was applied to maintain alignment with the instruction-tuned architecture of the base model. Label masking was enabled to ensure that loss was computed only over assistant response tokens, preventing unintended optimization over user prompts. Sequences were truncated to a maximum length of 2,048 tokens, with 98.7% of samples naturally fitting within this limit, indicating minimal truncation impact on training quality.

Training was performed using Paged AdamW (32-bit) optimizer with a learning rate of 1×10^{-4}. A cosine learning rate schedule was applied, including 50 warmup steps to stabilize early training dynamics. Weight decay was set to 0.01, and gradient clipping at 0.3 was applied to prevent instability during backpropagation. Mixed precision training was conducted using bfloat16 computation to balance numerical stability and memory efficiency.

Due to memory constraints and to reflect deployment-oriented conditions, a batch size of 2 was used. Gradient accumulation of 16 steps resulted in an effective batch size of 32. Training was conducted for 5 epochs, totaling 165 optimization steps. Model evaluation was performed every 50 steps, and early stopping with a patience of 3 evaluations was applied to prevent overfitting. Training was executed on a single NVIDIA A100 GPU, requiring approximately 43 minutes to complete.

Table 2. Training Configuration and Outcomes

Category	Parameter	Value
Preprocessing	Chat Template	Gemma-2 template
	Label Masking	Enabled
	Max Sequence Length	2048 tokens
	Natural Fit Rate	98.7%
Optimization	Optimizer	Paged AdamW (32-bit)
	Learning Rate	1×10^{-4}
	Scheduler	Cosine
	Warmup Steps	50
	Gradient Clipping	0.3
	Weight Decay	0.01
Training Setup	Batch Size	2
	Gradient Accumulation	16
	Effective Batch Size	32
	Epochs	5
	Total Steps	165
	Mixed Precision	bfloat16
	Evaluation Interval	Every 50 steps
	Early Stopping Patience	3
Hardware	GPU	NVIDIA A100
	Training Time	43 minutes
Outcome	Final Validation Loss	1.4775
	Perplexity	4.38
	Improvement vs Base	44.7%

As reported in Table 2, the final validation loss achieved was 1.4775, corresponding to a perplexity of 4.38. Compared to the base model configuration, this represents a 44.7% improvement in validation performance, indicating successful domain adaptation under parameter-efficient constraints.

3.5. Evaluation

To assess the effectiveness of the proposed domain-

adapted model, evaluation was conducted using three complementary perspectives: automatic text generation metrics, domain-specific reliability measures, and system-level performance analysis. Automatic evaluation metrics were used to quantify lexical and semantic similarity between generated responses and reference answers. ROUGE[18] measures token-level overlap, including unigram (ROUGE-1), bigram (ROUGE-2), and longest common subsequence (ROUGE-L) scores. BLEU[19] evaluates n-gram precision with a brevity penalty mechanism. METEOR[20] incorporates synonym matching, stemming, and word order considerations. BERTScore[21] computes contextual semantic similarity using pretrained multilingual BERT embeddings. These metrics provide quantitative indicators of response alignment with expert-written references.

Given the safety-critical nature of elevator emergency guidance, automatic metrics alone are insufficient. Therefore, domain-specific evaluation criteria were introduced. Faithfulness measures factual consistency with established safety procedures. A subset of 100 randomly selected test samples was annotated by certified professionals to assess technical accuracy. Answer Correctness evaluates whether responses directly address the user query and provide actionable guidance. Hallucination Rate identifies uncertain expressions (e.g., speculative language), fabricated specifications, or contradictory instructions that may compromise safety.

System-level performance was also evaluated. Latency was measured as average generation time per response, while throughput was calculated as the number of queries processed per second. Token statistics were analyzed to assess generation efficiency. All evaluations were conducted on an NVIDIA A100 GPU using identical generation parameters: temperature 0.1, top-p 0.95, and maximum output length of 256 tokens.

4. Results

4.1. Training and Quality Evaluation

Fine-tuning converged smoothly over five epochs, achieving a final validation loss of 1.4775 (perplexity 4.38), compared to the base model's perplexity of 7.92, indicating substantial domain adaptation. The comprehensive evaluation comparing the base and fine-tuned models is presented in Table 3.

Table 3. Comprehensive Evaluation: Base vs. Fine-Tuned Model

Metric	Base	Fine-Tuned	Abs. Δ	Rel. Δ (%)
ROUGE-1	0.2586	0.3168	+0.0582	+22.49
ROUGE-2	0.0716	0.1106	+0.0389	+54.36
ROUGE-L	0.1525	0.1986	+0.0461	+30.19
BLEU	0.0131	0.0417	+0.0286	+218.11
METEOR	0.1542	0.1759	+0.0217	+14.06
BERTScore F1	0.6559	0.7050	+0.0491	+7.49
Correctness (%)	54.49	60.49	+6.00	+11.01
Hallucination (%)	54.00	22.00	-32.00	-59.26
Avg Latency (s)	19.72	10.05	-9.67	-49.04
Throughput (s/s)	0.051	0.099	+0.049	+96.24
Avg Output Tokens	238.95	121.51	-117.44	-49.15

As shown in Table 3, the fine-tuned model consistently outperformed the base model across all automatic evaluation metrics. ROUGE-1 improved by 22.49%, indicating stronger lexical alignment with domain-specific terminology. ROUGE-2 showed a significant improvement of 54.36%, reflecting enhanced bigram precision and better phrase-level structuring in emergency instructions. ROUGE-L increased by 30.19%, suggesting improved sequence-level coherence.

BLEU exhibited the largest relative improvement (+218.11%), highlighting substantial gains in n-gram precision. Although BLEU scores remain numerically lower than ROUGE values, as typical in instruction-based generation tasks, the relative gain demonstrates stronger alignment with reference safety responses. METEOR improved by 14.06%, and BERTScore F1 increased by 7.49%, indicating enhanced semantic similarity and contextual alignment.

Beyond automatic metrics, domain-specific evaluation revealed critical improvements in procedural reliability. Answer correctness increased from 54.49% to 60.49%, while hallucination rate decreased dramatically from 54% to 22%, representing a 59.26% reduction. This reduction is particularly significant given the safety-critical context of elevator emergency guidance. The base model frequently produced uncertain or speculative language, whereas the fine-tuned model generated more grounded and directive responses aligned with established safety procedures.

Fine-tuning improved not only response quality but also system efficiency. Average latency decreased from 19.72 seconds to 10.05 seconds (−49.04%), largely due to shorter and more focused outputs. The average output length was reduced from 238.95 tokens to 121.51 tokens, reflecting improved response conciseness. Throughput nearly doubled, increasing by 96.24%, demonstrating that domain adaptation enhanced both computational efficiency and response clarity.

These results indicate that parameter-efficient domain-specific fine-tuning successfully improved lexical

alignment, semantic consistency, procedural correctness, and runtime performance simultaneously, an important requirement for embedded safety-critical applications. The radar visualization in Figure 2 provides a consolidated view of the performance differences between the base and fine-tuned models. The expanded area covered by the fine-tuned model indicates consistent improvements across lexical, semantic, reliability, and efficiency metrics. This visual summary reinforces the quantitative findings reported in Table 3.

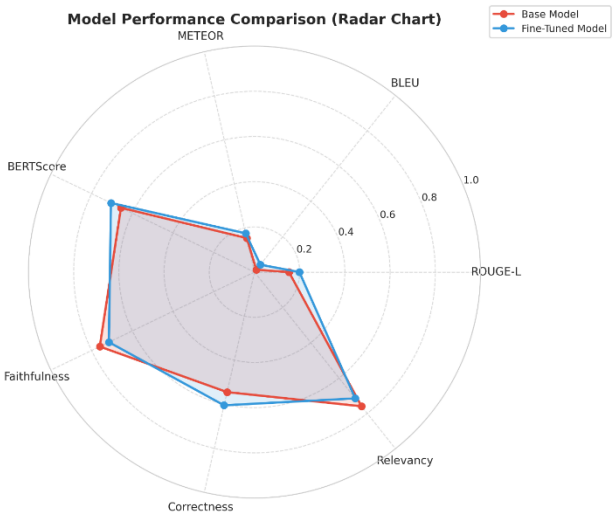

Figure 2. Radar chart comparing models. Fine-tuned model (blue) shows consistent improvements across metrics.

The radar chart presented in Figure 2 provides a multidimensional view of the performance differences between the base and fine-tuned models across six key evaluation dimensions: BLEU, ROUGE-L, Correctness, Relevancy, Faithfulness, and BERTScore. The fine-tuned model consistently occupies a larger area across all axes, confirming that improvements are not confined to a single dimension but are distributed uniformly across lexical, semantic, and reliability-oriented criteria. Notably, the most pronounced expansions are observed in the Faithfulness and Correctness axes, reflecting the model's improved alignment with established safety procedures and its ability to provide procedurally accurate responses. The BERTScore and ROUGE-L axes also demonstrate meaningful gains, indicating stronger contextual coherence and sequence-level alignment with reference responses. In contrast, the base model exhibits a considerably smaller and uneven coverage across the radar dimensions, suggesting inconsistent performance that is particularly inadequate for safety-critical deployment scenarios. Taken together, the radar visualization reinforces the quantitative findings reported in Table 3 and highlights that domain-specific fine-tuning produces a balanced and reliable model suitable for embedded emergency guidance applications.

4.2. Qualitative Example

To further illustrate the behavioral differences between the base and fine-tuned models, a representative emergency query was evaluated.

Query (Turkish): “Asansörde sıkıştım, kapı açılmıyor, ne yapmalıyım?” (“I am stuck in the elevator, the door will not open, what should I do?”)

Base Model Response (Turkish): “Öncelikle sakin olmaya çalışın. Asansörlerde genellikle acil durum butonları bulunur. Yardım istemek için bu butonları kullanabilirsiniz. Kapıyı zorlamamanız önerilir.” (“First, try to remain calm. Elevators generally have emergency buttons. You can use these buttons to request help. It is recommended not to force the door.”) The base model response is relatively generic and includes hedging expressions such as “genellikle” (“generally”). The instructions lack procedural sequencing and do not provide specific details such as button identification or expected rescue timeframes. The tone is cautious but not directive.

Fine-Tuned Model Response (Turkish): “Sakin kalın ve panik yapmayın. Alarm butonuna (kırmızı veya zil işaretli) basarak yardım çağırın. Kapıyı zorlamayın, tehlikelidir. Kabin havalandırması aktiftir. 112’yi arayın. Yardım 15–30 dakika içinde gelir. Oturup bekleyin.” (“Remain calm and do not panic. Press the alarm button (red or marked with a bell symbol) to call for assistance. Do not force the door, it is dangerous. Cabin ventilation is active. Call emergency services (112). Help will arrive within 15–30 minutes. Sit down and wait.”) The fine-tuned model produces a structured and sequential response with explicit procedural guidance. The instructions are directive, include identifiable physical cues (button appearance), and provide a time expectation for rescue. The language avoids uncertainty and maintains a confident tone, which is essential in crisis communication scenarios.

The qualitative comparison demonstrates that domain-specific fine-tuning substantially improves structural clarity, procedural specificity, and confidence calibration. While the base model produces a generally relevant response, it relies on generic phrasing and hedging expressions, resulting in less decisive guidance. In contrast, the fine-tuned model delivers instructions in a clear sequential order, ensuring that users can easily follow step-by-step actions during a stressful situation. The inclusion of actionable details, such as identifying the alarm button by its color or symbol and providing an estimated rescue timeframe, reflects the internalization of procedural knowledge during fine-tuning. Furthermore, the elimination of speculative language enhances decisiveness, which is particularly critical in emergency communication. These differences indicate that parameter-efficient domain adaptation not only improves lexical alignment and metric-based performance but also meaningfully enhances response reliability and practical usability in safety-critical elevator scenarios.

5. Discussion

The findings of this study demonstrate that parameter-efficient fine-tuning via QLoRA enables effective adaptation of large language models to safety-critical elevator emergency scenarios under embedded hardware constraints. As illustrated in Figure 3, the fine-tuned model consistently outperforms the base model across all automatic evaluation metrics, including ROUGE, BLEU, METEOR, and BERTScore. The visual comparison confirms that performance gains are not isolated to a single metric but reflect balanced improvements in lexical alignment, phrase-level structuring, and semantic similarity. This pattern indicates that domain-specific fine-tuning modifies not only vocabulary usage but also response structure and contextual coherence.

Also, findings are consistent with prior domain adaptation studies in other specialized fields. Similar to BioBERT and LEGAL-BERT, which demonstrated that embedding domain-specific knowledge into model parameters substantially improves task performance over general-purpose models, our results confirm that this principle extends to safety-critical industrial applications in low-resource language settings. Furthermore, the observed reduction in hallucination rate aligns with findings reported in elevator maintenance LLM studies, where factual reliability was identified as a primary concern in industrial deployments. The simultaneous gains in both response quality and inference efficiency also complement recent observations in parameter-efficient fine-tuning literature, suggesting that QLoRA-based adaptation represents a broadly viable strategy for resource-constrained domain specialization beyond the specific scenario investigated here.

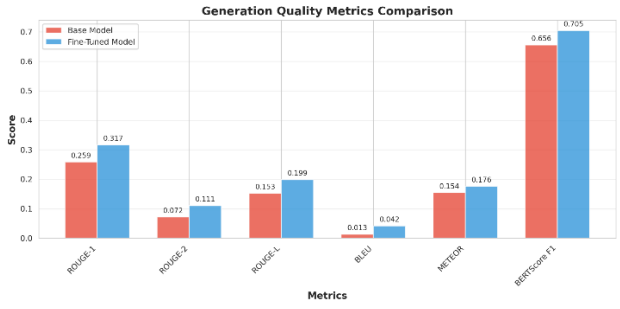

Figure 3. Generation quality metrics showing consistent gains across ROUGE, BLEU, METEOR, and BERTScore.

Beyond lexical and semantic gains, the most significant improvement concerns reliability. The substantial reduction in hallucination rate suggests that domain adaptation constrains generative variability and promotes more deterministic, grounded responses. In safety-critical environments such as elevator cabins, speculative phrasing or uncertain language can directly influence user behavior during emergencies. The fine-tuned model demonstrates clearer sequential structuring, directive phrasing, and confidence calibration, reflecting internalization of procedural emergency knowledge. These qualitative shifts align with the quantitative improvements observed in Table 3 and visually summarized in Figure 3.

Another notable outcome is the simultaneous enhancement of efficiency. The fine-tuned model generates shorter, more focused responses, resulting in reduced latency and improved throughput. This is particularly important for edge AI deployment, where computational resources are limited and response time is critical. Unlike many adaptation strategies that improve accuracy at the cost of increased computation, the proposed configuration improves both quality and runtime performance, making it well suited for embedded emergency systems.

From a system architecture perspective, these results support the selection of fine-tuning over retrieval-augmented generation (RAG) for this deployment scenario. While RAG offers flexibility in updating knowledge dynamically, it introduces reliance on network connectivity and additional latency. Elevator cabins (especially during emergencies) may lack stable connectivity, and deterministic offline operation becomes mandatory. Embedding procedural knowledge directly into model parameters through fine-tuning therefore provides a more robust and deployment-feasible solution for this context.

Despite the observed improvements, several limitations remain. The dataset, although expert-validated, may not fully cover rare or high-severity edge cases such as complex multi-person emergencies or cascading mechanical failures. Automatic evaluation metrics, while informative, cannot fully capture the appropriateness and contextual safety of responses. Furthermore, the residual hallucination rate indicates that human oversight mechanisms remain necessary. Techniques such as confidence thresholding, structured output templates, human-in-the-loop escalation for complex scenarios, and periodic expert auditing should be integrated into real-world deployments to ensure responsible use.

Although the dataset includes a number of short and fragmented queries reflecting realistic user behavior under stress, the construction process did not specifically prioritize panic language patterns such as broken sentences, single-word exclamations, or incoherent phrasing. Emergency situations frequently elicit such atypical linguistic inputs, and the model's robustness to these patterns was not systematically evaluated in the current study. Future work should incorporate a dedicated subset of panic-language samples into the training data, alongside targeted evaluation protocols designed to assess model performance under degraded input conditions.

Overall, the results suggest that quantization-aware, parameter-efficient fine-tuning provides a viable pathway for integrating LLM-based assistants into embedded safety-critical systems. By balancing procedural

reliability, semantic alignment, and computational efficiency, the proposed approach advances the practical application of large language models in real-world emergency response environments.

6. Conclusion

This study demonstrates that parameter-efficient fine-tuning using QLoRA enables effective adaptation of large language models to safety-critical edge AI applications. By fine-tuning Gemma2-9B-Instruct on a domain-specific Turkish elevator emergency dataset, substantial improvements were achieved across multiple dimensions, including generation quality, semantic alignment, procedural correctness, hallucination reduction, and runtime efficiency. Notably, these gains were obtained while updating only 37 million parameters (0.4% of the model) and maintaining a deployment-ready 4.8 GB quantized footprint.

The results confirm that domain-specific fine-tuning does more than improve lexical overlap; it reshapes response structure, confidence calibration, and procedural grounding, key attributes for emergency communication systems. The significant reduction in hallucination rate further underscores the importance of embedding domain knowledge directly into model parameters when operating in safety-critical contexts. From a deployment perspective, the proposed approach validates the feasibility of running large language models entirely offline on embedded hardware. Unlike retrieval-based architectures, the fine-tuned model operates deterministically without external network dependency, making it suitable for environments where connectivity cannot be guaranteed.

The practical implications of this work extend beyond the specific elevator emergency domain. The proposed pipeline (combining domain-specific dataset construction, parameter-efficient QLoRA fine-tuning, and 4-bit quantized deployment) provides a replicable framework applicable to a wide range of safety-critical edge AI scenarios, including industrial machinery operation, mining safety systems, and disaster response guidance. The demonstrated ability to reduce hallucination rates by 59.26% while simultaneously improving inference speed by 49% underscores that reliability and efficiency are not mutually exclusive objectives in embedded AI design. Furthermore, the relatively modest hardware requirements and short training time of 43 minutes on a single GPU suggest that this approach is accessible to organizations operating under real-world resource constraints, lowering the barrier for responsible AI deployment in industrial safety environments.

Overall, this work establishes a practical framework for integrating LLM-based assistants into constrained, real-world emergency systems. By combining quantization-aware fine-tuning, multi-metric evaluation, and safety-oriented design principles, the study contributes to advancing reliable edge AI solutions. Future research may explore continual learning for protocol updates, multimodal integration, multi-turn interaction optimization, and controlled user studies in simulated emergency environments to further enhance robustness and usability.

Acknowledgments

This research was supported by Butkon Elevator Industry Trade Inc., Research and Development Center. Also, the authors thank Hugging Face for open-source tools [22], the bitsandbytes library[23] for quantization, and elevator safety professionals whose validation contributed to dataset quality.

Declaration of Ethical Standards

This research does not involve direct interaction with human participants or animals. Accordingly, ethical approval and informed consent were not required.

Credit Authorship Contribution Statement

All authors contributed substantially and equally to this study. The processes of conceptualization, methodology development, software implementation, validation, formal analysis, investigation, data curation, manuscript drafting, review and editing, as well as visualization were carried out collaboratively. All authors have reviewed and approved the final version of the manuscript.

Declaration of Competing Interest

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Funding / Acknowledgements

This research received no external funding.

Data Availability

The dataset used in this study was constructed by the authors by synthesizing publicly available safety manuals, international elevator standards (EN 81, ASME A17.1), and academic literature. The structured question-answer pairs derived from these sources are available from the corresponding author upon reasonable request.

References

[1] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems,* vol. 30, 2017.

[2] T. Brown *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems,* vol. 33, pp. 1877-1901, 2020.

[3] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems,* vol. 35, pp. 27730-27744, 2022.

[4] H. Touvron *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971,* 2023.

[5] G. Team *et al.*, "Gemma 2: Improving open language models at a practical size," *arXiv preprint arXiv:2408.00118,* 2024.

[6] Z. Ji *et al.*, "Survey of hallucination in natural language

generation," *ACM computing surveys,* vol. 55, no. 12, pp. 1-38, 2023.

[7] A. Conneau *et al.*, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 8440-8451.

[8] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems,* vol. 33, pp. 9459-9474, 2020.

[9] E. J. Hu *et al.*, "Lora: Low-rank adaptation of large language models," *ICLR,* vol. 1, no. 2, p. 3, 2022.

[10] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," *Advances in neural information processing systems,* vol. 36, pp. 10088-10115, 2023.

[11] J. Lee *et al.*, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics,* vol. 36, no. 4, pp. 1234-1240, 2020.

[12] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The muppets straight out of law school," *arXiv preprint arXiv:2010.02559,* 2020.

[13] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," *arXiv preprint arXiv:1903.10676,* 2019.

[14] E. Alsentzer *et al.*, "Publicly available clinical BERT embeddings," in *Proceedings of the 2nd clinical natural language processing workshop*, 2019, pp. 72-78.

[15] D. Araci, "Finbert: Financial sentiment analysis with pre-trained language models," *arXiv preprint arXiv:1908.10063,* 2019.

[16] P. Zare, "Enhancing maintenance strategies for elevators through fine-tuning large language models," 2024.

[17] K. Chakma, "Evaluating enterprise product recommendation chatbot using LLM: the case of easy selection," 2025.

[18] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74-81.

[19] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311-318.

[20] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65-72.

[21] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675,* 2019.

[22] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, 2020, pp. 38-45.

[23] T. Dettmers, M. Lewis, S. Shleifer, and L. Zettlemoyer, "8-bit optimizers via block-wise quantization," *arXiv preprint arXiv:2110.02861,* 2021.