International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 12 December 2025
Page No: 4214-4217
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0617

A PEFT (Parameter Efficient Fine-Tuning) Framework Enabling Context Augmentation in LLMs

Navaneeth D[1], Mijaz Mukundan[2], Jishy Samuel[3]
[1]Student, International Institute of Information Technology, Electronic City Phase–1, Bengaluru, Karnataka, India.
[2,3]Scientist/Engineer, MSSG, Vikram Sarabhai Space Centre, Thiruvananthapuram, Kerala, India.
Emails: *navaneeth.d@iiitb.ac.in[1], mijaz_mukundan@vssc.gov.in[2], jishy_samuel@vssc.gov.in[3]*

Abstract
This paper explores domain-specific fine-tuning of large language models (LLMs) to enhance their contextual understanding in specialised tasks. Mistral-7B was selected as the base model due to its open-source availability, efficiency, and performance. Training was conducted using PDF-derived datasets, converted into structured formats suitable for model ingestion. The parameter-efficient fine-tuning technique LoRA was employed, allowing training on consumer-grade hardware. The fine-tuned models were deployed in a Retrieval-Augmented Generation (RAG) based chatbot system. Evaluation was carried out through human judgment, comparing outputs from the fine-tuned and base models using identical prompts. Results indicated that while the trained models effectively learned specific domain knowledge and keywords, they sometimes lacked deeper conceptual linkage between ideas. Nonetheless, the approach demonstrated a practical path toward resource-efficient customisation of LLMs for domain-specific applications.

Keywords: *Large Language Models (LLMs); LoRA; Retrieval-Augmented Generation*

1. Introduction
Large Language Models (LLMs) have transformed natural language processing through their impressive capabilities in language generation, summarization, and question answering. However, their effectiveness often diminishes when applied to domain-specific tasks requiring nuanced contextual understanding or reasoning beyond general-purpose knowledge. This is especially critical in applications where factual accuracy and domain alignment are essential. Although pretrained LLMs like Mistral-7B [1] are powerful, they are trained on general datasets and often lack the depth needed for specialized fields. Moreover, since fine-tuning the entire model is computationally expensive and not always feasible, especially under hardware constraints, an efficient solution is needed. The motivation for this project stems from the need to build a reasoning engine based on an LLM, but within the restriction of using only pretrained models. To address the contextual limitations of these models, the project explores Parameter-Efficient Fine-Tuning (PEFT) techniques that allow customization of LLMs for specific domains without modifying the entire model. By using LoRA (Low-Rank Adaptation) [2], this study aims to augment the reasoning capabilities of Mistral-7B through low-resource fine-tuning, making domain-specific adaptation accessible even on consumer-grade hardware. This approach not only lowers the computational barrier but also enables faster iterations and modular deployment, highlighting a practical path to building more intelligent, domain-aware LLM-based systems.

1.1. Objectives
This project focuses on enriching large language models (LLMs) with domain-specific knowledge by fine-tuning them on curated datasets. The objective is to enhance the model's ability to understand and respond to context-specific prompts more accurately. To achieve this, the parameter-efficient fine-tuning method LoRA was employed, which significantly

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 12 December 2025
Page No: 4214-4217
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0617

reduce computational requirements while maintaining performance.

- To extract and preprocess data from PDF documents and convert it into a suitable format for LLM training.
- To fine-tune large language models (LLMs) using parameter-efficient method LoRA.

2. Background

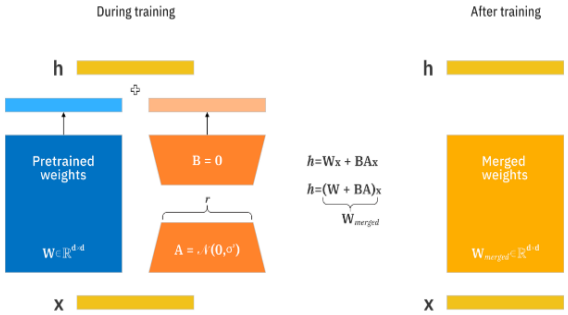

Figure 1 Understanding LoRA [3]

LoRA is an efficient and compact fine-tuning method that greatly minimizes the number of parameters requiring training. It does this by adding a small set of additional weights to the model and updating only these during training. As a result, LoRA enables much faster and more memory-efficient training, producing lightweight model files—typically just a few hundred megabytes—that are simple to save and distribute. LoRA (Low-Rank Adaptation) fine-tunes large models efficiently by freezing the original model weights and adding a lightweight, trainable component. Specifically, it inserts two low-rank matrices A and B into each layer of the model, forming an update of the form $W + BA$ without modifying the original weight matrix W. During training, only these smaller matrices are updated using gradient descent. This significantly reduces the number of trainable parameters and memory usage. After training, the product BA is merged into the original weights to create an updated model, keeping the base model intact. LoRA relies on the insight that many changes needed for adaptation lie in a low-dimensional space, so even small-rank updates can be effective. [3] This concept is further illustrated using a matrix example (Fig. 2): instead of learning a full 4×4 weight matrix (16 parameters), LoRA decomposes it into two matrices of size 4×1 and 1×4, using only 8 parameters for the same effect, shown in Figure 1 & 2.

3. Method

The base model selected for fine-tuning was Mistral-7B, an open-source LLM developed by Mistral AI. Training was initially conducted using Mistral-Finetune [4], the official tool provided by Mistral AI.

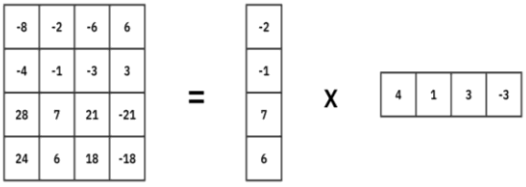

Figure 2 An Example Showing How LoRA Decomposes the Matrices [3]

The dataset used for training was constructed by extracting text from a collection of domain-relevant PDF documents. This textual data was cleaned and converted into a ".jsonl" format suitable for LLM training. After fine-tuning, the models were integrated into a Retrieval-Augmented Generation (RAG)-based chatbot system. To evaluate the effectiveness of the domain-specific training, both the fine-tuned and original (untrained) versions of the models were queried using identical prompts. Their responses were then assessed by human evaluators with expertise in the target domain. The evaluation revealed that the fine-tuned models successfully learned relevant terminology and context-specific information. However, they occasionally exhibited difficulties in establishing logical connections between related concepts unless explicitly prompted. This highlighted the limitations of lightweight fine-tuning in terms of semantic reasoning, while still demonstrating significant gains in factual recall and relevance within the target domain.

4. Training

4.1. Extracting Datasets

In this study, the dataset was obtained from the textual content extracted from multiple PDF. This

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 12 December 2025
Page No: 4214-4217
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0617

process involved parsing and preprocessing the PDFs to accurately capture and structure the textual content for analysis. This method ensured the systematic collection of relevant information while maintaining the integrity and reliability of the original data.

4.2. Selecting Model

Mistral-7B was selected for this work due to its strong balance between performance and efficiency. With 7 billion parameters and a dense transformer architecture, it supports both causal language modeling and instruction-following tasks. Its optimization for low-latency inference and sliding window attention enables effective handling of long contexts with minimal computational cost, making it a competitive and practical choice—especially when fine-tuned with methods such as LoRA.

4.3. Tool For Training

The Mistral-Finetune repository was selected for this work due to its efficiency and flexibility in fine-tuning Mistral models. It enables parameter-efficient adaptation through LoRA, allowing effective customization on domain-specific datasets without altering the base model weights. The repository's support for multiple data formats (e.g., Alpaca, ShareGPT, etc.) and compatibility with both Mistral-7B and Mixtral [5] models make it a practical and resource-efficient choice for fine-tuning while maintaining model stability and performance.

4.4. Merging The Weights

Mistral-finetune allows both storing the LoRA-trained weights and the model merged with the LoRa-trained weights. To combine the LoRA weights to the base model, we might need to use a script provided by the HuggingFace [6]. After that the weights can be merged using the transformers library in python.

4.5. Deployment in Ollama

In order to facilitate seamless local deployment of large language models, Ollama [7] was selected for its simplicity, efficiency, and ability. It enables the execution and management of models such as Mistral directly on local hardware with minimal configuration, ensuring data privacy and low-latency inference. Ollama requires models to be in the GGUF format [8]; therefore, the fine-tuned model was converted into this format using the conversion script provided by "*Llama.cpp"* [9]. Once converted, the model was ready for deployment within the Ollama environment.

5. Results

The evaluation was conducted by comparing the responses of the base model and the fine-tuned model to identical prompts related to the IRNSS-1H [10] mission. The inputs and outputs produced by the base and fine-tuned models are presented in Figures 3, 4, 5, and 6**.** The base model produced brief and generic answers with limited technical accuracy (see Figures 3 and 4). In contrast, the LoRA fine-tuned model demonstrated a more detailed and structured understanding of the mission (see Figures 5 and 6). It correctly explained that the IRNSS-1H satellite was launched aboard PSLV-C39 to replace the failed IRNSS-1A, and that the mission failed due to the payload fairing (heat shield) not separating during flight. This malfunction caused the satellite to remain trapped inside the fairing, preventing it from reaching its intended orbit. These details were verified against official data, including ISRO releases and publicly available records [11], confirming the accuracy of the fine-tuned model's responses. The base model failed to mention key technical aspects such as the heat shield malfunction and its effect on orbital deployment. While the fine-tuned model successfully reflected domain-specific knowledge and terminology, it occasionally struggled with linking related concepts across prompts. Nevertheless, the results strongly suggest that parameter-efficient fine-tuning using LoRA significantly improves the domain relevance and factual accuracy of large language models.

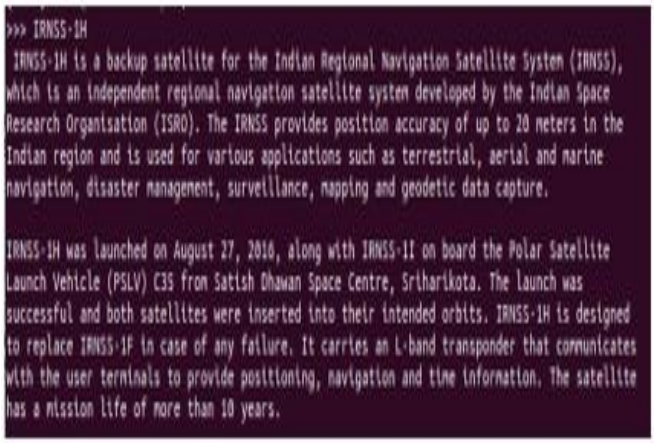

Figure 3 Output Obtained from Untrained Model

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 12 December 2025
Page No: 4214-4217
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0617

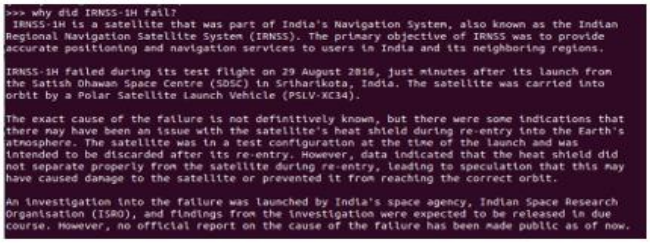

Figure 4 Output Obtained from Untrained Model

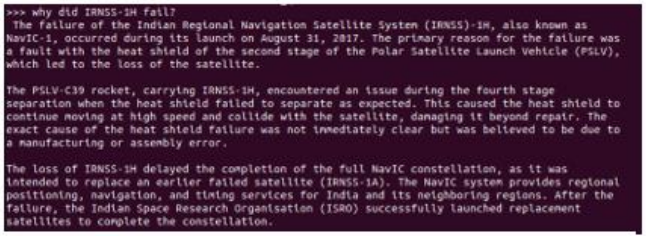

Figure 5 Output Obtained from LoRA-Trained Model

Figure 6 Output Obtained from LoRA-Trained Model

Conclusion

This project demonstrates the effectiveness of customizing large language models using parameter-efficient fine-tuning techniques like LoRA. It enabled adapting models such as Mistral-7B to specific domains with modest computational resources, without sacrificing performance. From setting up the training pipeline to deploying inference-ready models via Ollama, the workflow is accessible and reproducible, supporting adoption in low-resource environments. Comparative evaluations against the base model show substantial improvements in factual accuracy and domain awareness. While challenges remain—especially in semantic reasoning and generalization—this work provides a strong foundation for future enhancements, including multimodal integration, scalable adapter systems, and automated evaluation tools. Overall, it showcases the feasibility of domain-specific LLM customization and paves the way for democratizing AI in specialized fields requiring high accuracy and contextual understanding.

References

[1]. Mistral AI Team, “Mistral 7B | Mistral AI,” Mistral AI, 27 September 2023. [Online]. Available: https://mistral.ai/news/announcing-mistral-7b.
[2]. Hugging Face, “LoRA,” Hugging Face, [Online]. Available: https://huggingface.co/docs/diffusers/main/en/training/lora.
[3]. J. Noble, “What is LoRA (low-rank adaption)?,” IBM Corporation, 24 October 2023. [Online]. Available: https://www.ibm.com/think/topics/lora.
[4]. Mistral AI, “Mistral-Finetune,” Mistral AI, [Online]. Available: https://github.com/mistralai/mistral-finetune.
[5]. Hugging Face, “Mixtral,” Hugging Face, [Online]. Available: https://huggingface.co/docs/transformers/en/model_doc/mixtral.
[6]. Hugging Face, “convert_mistral_weights_to_hf.py · main · Hugging Face / Transformers,” Hugging Face, [Online]. Available: https://github.com/huggingface/transformers/blob/main/src/transformers/models/mistral/convert_mistral_weights_to_hf.py.
[7]. Ollama, “Ollama,” Ollama, [Online]. Available: https://ollama.com/.
[8]. Hugging Face, “GGUF,” Hugging Face, [Online]. Available: https://huggingface.co/docs/hub/en/gguf.
[9]. ggml-org, “llama.cpp · ggml-org/llama.cpp,” ggml-org, 10 March 2023. [Online]. Available: https://github.com/ggml-org/llama.cpp.
[10]. Indian Space Research Organisation (ISRO), “IRNSS-1H,” Indian Space Research Organisation (ISRO), [Online]. Available: https://www.isro.gov.in/IRNSS_1H.html.
[11]. Wikimedia Foundation, “IRNSS-1H - Wikipedia,” Wikimedia Foundation, 20 July 2025. [Online]. Available: https://en.wikipedia.org/wiki/IRNSS-1H.