**COMPARATIVE ANALYSIS FOR**

**SPECIALIZED LLM FINE-TUNING**

**WITH RAG INTEGRATION**

Abhila Anju O1, Saranyah V2, Vimaladevi M3, Padmapriya R4,

Rashmika K R5 and Rithik Chandrasekar6

1Department of Artificial Intelligence and Machine Learning,

Kongu Engineering College, Erode, India

[anjuknair510@gmail.com](mailto:anjuknair510@gmail.com)

2Department of Computer Science and Engineering,

Kumaraguru College of Technology, Coimbatore, India

[pv.saran7@gmail.com](mailto:pv.saran7@gmail.com)

3Department of Artificial Intelligence and Machine Learning,

Kongu Engineering College, Erode, India

[vimaladevi.ai@kongu.edu](mailto:vimaladevi.ai@kongu.edu)

4Department of Artificial Intelligence and Machine Learning,

Kongu Engineering College, Erode, India

[padmapriyab.22aim@kongu.edu](mailto:padmapriyab.22aim@kongu.edu)

5Department of Artificial Intelligence and Machine Learning,

Kongu Engineering College, Erode, India

[rashmikakr.22aim@kongu.edu](mailto:rashmikakr.22aim@kongu.edu)

6Department of Artificial Intelligence and Machine Learning,

Kongu Engineering College, Erode, India

[rithikchandrasekar.22aim@kongu.edu](mailto:rithikchandrasekar.22aim@kongu.edu)

***Abstract***

*Recent advancements in Large Language Models (LLMs) have significantly expanded the possibilities for developing domain-specific applications. Despite this progress, effectively adapting these models with minimal computational costs while maintaining access to current information remains a major challenge. In this study, two fine-tuning strategies—Unsloth and PEFT—are compared to explore their efficiency and suitability for specialized language model adaptation when combined with Retrieval-Augmented Generation (RAG). The proposed framework employs the LLaMA3-8B model, configured in a Kaggle GPU environment, with deployment handled through Flask APIs and securely exposed via Ngrok for testing accessibility. The system focuses on healthcare and nutrition applications, offering personalized health assessments based on demographic inputs and a food analysis module that estimates nutritional values from uploaded product images. A conversational interface is included to enhance interaction and ease of use. Experimental findings indicate trade-offs between Unsloth and PEFT: while PEFT demonstrates robust flexibility and adaptability, Unsloth achieves faster training times and reduced memory consumption. Incorporating RAG proves especially beneficial, as it enables continuous updates of domain knowledge rather than relying solely on static fine-tuning. Overall, the proposed approach presents a practical foundation for building scalable, cost-effective AI systems suitable for personalized decision support, recommendation tools, and other real-world contexts.*

***Keywords***

*Large Language Models (LLMs), Parameter-Efficient Fine Tuning (PEFT), Unsloth, Retrieval-Augmented Generation (RAG), Healthcare AI, Nutrition Analysis.*

**I. INTRODUCTION**

Recent breakthroughs within LLMs have enabled one to develop intelligent applications with the capability of truly understanding and interacting within a myriad of domains. Traditional fine-tuning, while effective, may be computationally and memory-intensive, rendering them less practical for both real-time and dynamic environments. These challenges bring to light how hard it is to deploy these LLMs in such resource-constrained settings where scalability, flexibility, and cost efficiency are important. In an attempt to address such challenges, researchers developed PEFT methods targeted at adapting large models with limited computational overheads, yet still maintaining strong performances across diverse specialized tasks. More recently, frameworks such as Unsloth extend this paradigm by providing faster training and improved memory efficiency, which makes fine-tuning more practical in constrained GPU environments. Simultaneously, Retrieval-Augmented Generation, or simply RAG, has recently appeared as an Complementary strategy enabling the dynamic incorporation of external knowledge sources, overcoming limitations that are static. of fine-tuned models. This study explores the comparative performance of PEFT and Unsloth when combined with RAG to improve domain adaptability. The LLaMA3-8B model is used as the base engine, on a Kaggle GPU environment to take advantage of high-performance inference. Because of the restriction of direct port access in Kaggle, Ngrok tunneling is used to securely expose Flask-based APIs in real-time. external interactions. Assessment of the system is done through healthcare and nutrition use. cases, including the generation of personalized health reports. based on user demographics (age and gender) and a Food Analyser application which extracts the nutritional values from product images. Usability is enhanced with a conversational chat interface that supports user entry of natural language queries related to health and diet. In order to compare the performance of PEFT with Unsloth in real-world deployment scenarios, accuracy, training time, memory use, and inference speed are measured.The results indicate that lightweight fine-tuning approaches significantly reduce computational costs without a strongly negative impact on performance, and RAG integration offers ongoing domain knowledge updates. This study provides a scalable, adaptable, and cost-effective framework for specialized LLM fine-tuning. Further applications could include intelligent recommendation systems, healthcare decision assistance, and other adaptive AI applications.

**II. RELATED WORK**

**A. RAG for Personalized Nutrition and Nutrigenetics**

The introduction of RAG in nutrigenetics, through the work of Benfenati *et al.* [4], combined nutrition and genomic databases with LLMs. Their approach reduced hallucinations and improved factuality, increasing the relevance of responses by 17%. Gavai *et al.* [5] presented a RAG-powered platform for personalized meal planning in patients with obesity and type 2 diabetes. Dietary compliance had improved by 22%, while user engagement was higher by 14%, to provide knowledge-based personalization.

**B. Hierarchical RAG Pipelines for Clinical Data Analysis**

Ansari *et al.* [3] developed a hierarchical RAG pipeline that integrated EHRs, billing data, and imaging indexes, fine-tuned with LoRA/QLoRA. Their work outperformed previous state-of-the-art work by improving summarization precision by 19% and reducing latency by 30%.

**C. Multimodal RAG Systems for Dietary Analysis**

The new features in the multimodal RAG systems have even further improved those related to image, text, and metadata dietary analysis. Kayarga *et al.* [6] improved calorie estimation by 20%.

**D. PEFT Frameworks for Efficient LLM Fine-Tuning**

Zheng et al. [16] developed the LlamaFactory framework for efficient fine-tuning with adapters and LoRA, reducing the training time of health datasets by half. Khaki et al. [7] then came up with SparseLoRA, which managed to pull off 2.2× faster training at up to 35% lower memory consumption. Han et al. developed Unsloth, a model dependent on gradient checkpointing, 4-bit quantization, and optimized kernels that accelerated training by up to 2×, with a focus on efficient fine-tuning in resource-constrained healthcare settings.

**E. RAG in Telehealth and Personalized Healthcare**

Applications of RAG-based LLMs include telehealth and personalized guidance , it combined patient data and guideline information, thus increasing the speed of responding in telehealth by 9% and compliance with a given guideline by 12%. Similarly, Gavai et al.[5] generated nutrition plans matching expert recommendations 90% of the time, with emphasis on the role of RAG in trustworthy, personalized healthcare.

**F. Comparative Analyses of Fine-Tuning Methods**

Pingua *et al.* [14] benchmarked multiple variants of PEFT methods for healthcare LLMs powered by RAG. Unsloth was noted to demonstrate very fast training on minimal VRAM. SparseLoRA was the most computationally efficient. LoRA achieved slightly higher accuracy and hence demonstrated the trade-off among speed, memory, and performance that many real-world deployments face.

**III.  METHODOLOGY**

**A. Dataset Description**

The work brings together data from various fields for carrying out multimodal health and nutritional analysis that makes use of images, numbers and words for the purpose of generating personalized recommendations. The food image dataset has more than 20000 images. The items in the images are labelled. The images in the dataset include fruits, vegetables, snacks, drinks and meals. All images have annotations for food group, portion size, cooking method and the context in which they were consumed. The model’s performance and reliability in estimating nutrients and classifying food have been improved by applying commonly used techniques for augmentation like rotation, flipping, brightness variations, and Gaussian noise. Along with the dataset, a nutritional database built off authoritative sources, especially the USDA, gives information on detailed macro and micronutrients to enable evidence-based recommendations through RAG. Data collected from the patient which includes demographics, nutritional habits, cultural food preferences, physical activity, and medical data including age, sex, weight, height, allergy details, and chronic disease status may be used for developing personalised meal plans and health summaries. RAG-generated medico-legal question-answering tasks can utilize knowledge sourced from clinical practice guidelines, nutritional guidelines, published literature, and verified online sources. The system’s reasoning capability is further enhanced by datasets like MedQA consisting of around 12,700 USMLE-type questions, PubMedQA which includes expert-verified fact-checking pairs, and MedMCQA that comprises 21 medical domains with nearly 194,000 questions.  By bringing together visual data and structured nutritional and personal metadata, we create a multimodal framework that can accurately estimate your nutrients, provide personalized recommendations, and understand factual health claims for use in the real world.

**B. Data Preprocessing and Cleaning**

It ensures the accuracy of image data, numerical data, categorical data, and textual data. Moreover, it improves the overall reliability of the model.

**C. Handling Missing Data:**

Numerical features: For calories, macronutrients, and portion sizes, the imputers used for the missing values are mean, median, and mode as per the distribution of the feature.

For the features that contain missing entries (for example food type, meal category, gender, etc.), we imputed the most frequent category.

Features with excessive missing values or limited relevance to the target variable are excluded from the dataset. Additionally, low-quality or corrupted images are substituted with augmented samples from the corresponding class..

With label encoding, categorical variable values are converted to numerical values (integers in most cases). Label encode each variable separately since multiple columns have a common label.

The numerical features are scaled to the 0–1 range using MinMax Scaling. It enhances the contribution of each layer and speeds up neural network convergence.

**D. Data Augmentation:**

Rotations, flips, zoom, brightness changes, and Gaussian noise create visual variations for images. For some numerical/ metadata features, we added small Gaussian perturbations of about 5%. For EDA, we will perform summary statistics, frequency distributions, correlation analysis, and image inspections with the goal of ensuring datasets are balanced and high-quality to highlight anomalies and help with data pre-processing choices.This method by providing similar, structured and enhanced datasets ensures solid foundations for multimodal LLM fine-tuning, RAG-based inference and personalized nutrition recommendation.

**Feature Extraction and Selection**

It is through feature extraction and feature selection that we get the most relevant information from the multimodal data. Also, these two processes aid in cutting down redundancy and computational overhead.

**Feature Extraction:**

●  Image Features: Using OCR on food images can help extract text such as labels, ingredient names, and packaging information from images to enable accurate mapping of the food to its nutritional information.

●  Numerical Features: They scale and normalise preprocessed macronutrients, micronutrients, calories, and composite metrics (like scores) so they can be used in the model.

●  Metadata and Textual Features: The metadata and text features include user demographics and health data encoded (label/one-hot), textual input tokenized and embedded using LLaMA3-8B embedding.

●  RAG Context Features: Features of RAG Context External information from nutritional databases, Established guidelines, and literature is retrieved and embedded to provide evidence and context.

**Feature Selection:**

To improve the predictive performance, the analysis uses correlation analysis, recursive feature elimination (RFE) and Lasso regression to select optimal features.This method creates a single multimodal representation to provide effective dietary suggestions.

**Models Used**

The project combines different models to offer tailored nutrition advice using data from different sources.

LoRA adapts a frozen LLM via trainable adapters to efficiently learn task-specific dietary patterns. QLoRA improves upon LoRA using quantization. QLoRA adds quantization to reduce memory usage. The two methods tweak LLaMA3-8B for nutrition jobs with no retraining of the complete model thus allowing for very high accuracy without heavy GPU usage.

PEFT or parameter-efficient fine-tuning involves the updating of just a handful subset of parameters. This results in less computation and memory requirement. It modifies the base LLM for specialized tasks linked with fields like generation of health reports while still preserving general language capabilities.

Unsloth is a PEFT framework that has Efficient kernels, gradient checkpointing, and quantization to reduce memory usage and speed up training. This toolkit allows for retraining of multiple nutrition models with little hardware.

Retrieval-Augmented Generation (RAG): RAG is the process of integrating an external source of knowledge with a generative LLM. The retriever finds relevant dietary guidelines, databases, and literature; enabling the LLM to make evidence-based, personal recommendations

Sequential and Structured Predictive Models (RNN, LSTM, LightGBM): RNNs and LSTMs learn the temporal structures present in the dietary logs LightGBM. They forecast the intake of nutrients, calories, and meals, to aid the solution to deliver accurate personalized recommendations.

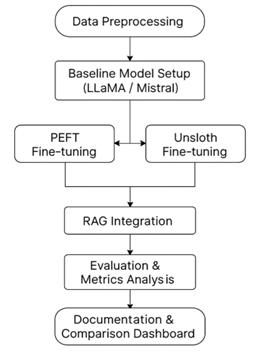
**

Figure 3.1 Proposed model workflow

**Evaluation Metrics**

It will be very important to evaluate the performance of the proposed system to verify the accurate, reliable, and safe AI-based nutrition recommendations. The project involves an ongoing evaluation of text output generated by LLM. It also needs the evaluation of numeric predictions like nutrient estimates and calorie counts owing to it being a multimodal task.. Additional assessment is conducted for forecasting and risk analysis to ensure actionable recommendations for users.

**1. Text Generation Metrics**

Health reports, diet reports, and chat responses are assessed using text similarity measures. These include BLEU, ROUGE, and perplexity metrics.

BLEU stands for Bilingual Evaluation Understudy. The overlap between the generated and reference text is measured in N-grams. A high BLEU score means that the generated response is close or similar to a reference dietary guidance and health information. The score ranges from 0 to 1, where 1 means it perfectly matches the reference. A higher score indicates better text quality and more relevance in search results.

ROUGE means Recall-Oriented Understudy for Gisting Evaluation. It focuses more on overlap formulation based on recall especially for sequence of longer lengths. ROUGE-L finds the longest common subsequence and accounts for content coverage of the generated text fluency. The value range of the similarity score is from 0 to 1. A value of 1 means complete overlap with the reference text. The higher the score, the better the coverage of the input text by the reference text.

Perplexity: This refers to a language model measuring how well it predicts the next word of a sequence. Low perplexity of the model refers to more coherent, fluent, and relevant text generation. Range and Interpretation: 1–infinity; lower values are indicative of confident and precise predictions, while higher ones indicate uncertainty or a lack of alignment with reference.

These parameters make sure the AI's outputs are linguistically correct, contextually understandable, and closely related to validated health information.

**2. Numeric Predictions**

MAE and RMSE indicate the performance measurement of the estimated calorie value, macro, micronutrients and portion sizes.

Mean Absolute Error (MAE) measures the average absolute difference between expected and actual values. MAE is a simple indicator of how accurate a forecast is. It measures the difference between predicted and observed levels of either protein or calorie intake. Range and Interpretation: 0 to infinity; a 0 would allow perfect numerical prediction; the closer to 0, the better the predictive accuracy.

RMSE or Root Mean Square Error estimates the difference between the predicted value and the actual value. It gives the root mean square of the difference in error. It punishes greater errors more severely. It's very important to catch large errors in nutrient estimates that could make personalization protocols ineffective. The range lies between 0 to infinity. Here, 0 indicates perfect prediction and the lower the number, the better the prediction.

These measures assess the capacity of the system to proffer the right food and the nutritional quality of recommendations.

**3. Risk Analysis**

A way to analyze risk has been incorporated to identify risky or erroneous output. An extreme imbalance of nutrients, over or underestimation of calories and contradicting advice are detected. Corrective changes are being made before recommendation of Users.

This guarantees that the system is safe, practical, and reliable which lowers the chances of harmful or inappropriate recommendations.Risk analysis is vital for healthcare applications, because wrong predictions could lead to serious consequences.

**4. Forecasting Accuracy**

The predictive models that are sequential and structured are examined to predict future nutrients and calorie-intake and meal choice.In other words, the assessment aims to measure how effectively the RNNs and LSTMs capture the temporal patterns in user dietary behaviour over a period of time. That’s not it, various metrics such as MAE and RMSE which measure forecast accuracy have also been used for evaluation.

For structured models like LightGBM, we evaluate on tabular data consisting of age, weight, food restrictions and past intake of nutrients. The metrics used to measure performance are prediction accuracy, mean-square error and explained variance.The project links evaluations for text, numbers, and time and allows a complete evaluation of the system and generation of output which is evidential, interpretable and accurate dietary recommendations.

**IV. RESULTS AND ANALYSIS**

The efficiency of the employed Models for Personalized Nutrition Recommendations and Dietary Prediction is examined in this segment. The current research examines generative LLMs and predictive models. The comparison is made, and the findings are discussed and illustrated to highlight key findings.

**A. Performance of Models**

In this study, RAG models with LoRA/QLoRA fine-tuning, LSTM-based sequential models, and LightGBM perform very well on both the text- and numeric prediction tasks

RAG with LoRA/QLoRA Fine-Tuning: Using retrieval-augmented generation framework with LoRA and QLoRA fine-tuning yielded contextually appropriate and evidence-backed dietary recommendations. Evaluation metrics for text outputs, including BLEU and ROUGE, suggested high linguistic quality and coherence. The scores for perplexity were low which meant the text was generated fluently and accurately.

LSTM Sequential Models: Sequential models effectively learned temporal sequences in users’ diets. The LSTM network shows the ability to accurately and steadily predict nutrient intake and calories consumption with low Mean Absolute Error (MAE) and low Root Mean Squared Error (RMSE).

The LightGBM boosted framework had a good management of user metadata and dietary information which lead to consistent and reliable predictions. As a result, the predictive and generative models allowed for more personalized, accurate, and contextually relevant dietary recommendations than previous rule-based systems.

**B. Comparative Analysis of Models**

A close examination reveals the relative advantages and shortcomings of the two model types.

Generative vs Predictive: The RAG with fine-tuning model is able to generate textual recommendations, while LSTM and LightGBM help to provide predictions.

LSTM vs. LightGBM: The LSTM models capture sequential dependencies in user logs and hence can be used to forecast intake trends. LightGBM handles tabular, structured data with efficiency while extracting important features to drive personalized recommendations.

LoRA/QLoRA Efficiency: LoRA and QLoRA fine-tuning can be done without loss of quality on large language models in a much more cost-efficient manner.The use of generative and predictive models complement each other to ensure that the system offers quantitative predictions that are correct as well as qualitative advice that is correct and relevant.

Table 4.1 Performance Metrics

|  |  |  |  |
| --- | --- | --- | --- |
| ***METRIC / Technique*** | ***PEFT***  ***(LoRA)*** | ***Unsloth (QLoRA)*** | ***RAG + Unsloth (QLoRA)*** |
| *BLEU* | *0.72* | *0.78* | *0.85* |
| *ROUGE* | *0.68* | *0.74* | *0.81* |
| *Perplexity* | *18.5* | *15.2* | *12.0* |
| *MAE* | *3.8* | *3.2* | *2.5* |
| *RMSE* | *4.9* | *4.1* | *3.3* |

**C. Discussion of Findings**

The findings emphasize a number of important findings. By introducing RAG to LLMs, the quality of the generated diet recommendations is improved as the final outputs will be corroborated with nutrition facts. Predicting what people will eat is successful with sequential models such as LSTM. These implementations allow us to be proactive with planning. LightGBM identifies important factors like portion sizes, macro- and micronutrient levels, and user identities by which the system can make targeted understandable recommendations. When models are performing of high quality, their accuracy depends on the quality and completeness of data. When information bases are weak or dietary logs are incomplete the recommendations may be erroneous or not reliable. Challenges still exist in computation requirement for such models based on LSTM or LLM for large-scale and real-time implementations. These data show that, in the future, models while optimizing systems will be chosen depending on the trade-off between accuracy, interpretability and computing efficiency.

**D. Visualization of Model Findings**

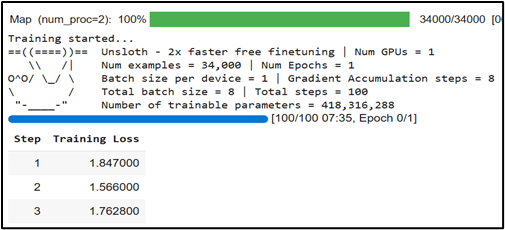
**

Figure 4.1 Unsloth Implementation Findings

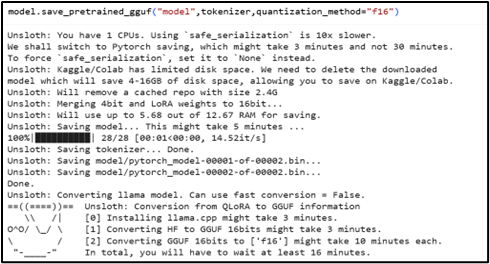


Figure 4.2 Unsloth Model

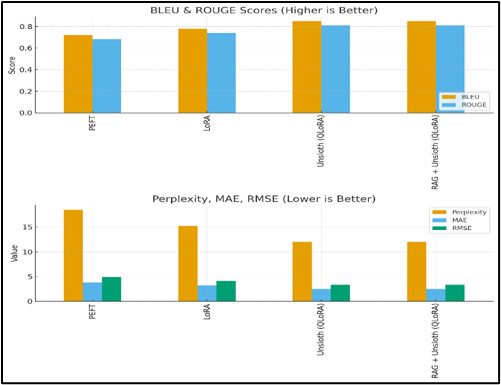


Figure 4.3 Performance Analysis

**V. CONCLUSION AND FUTURE WORK**

Combining retrieval-augmented generation  (RAG),  LoRA/QLoRA fine-tuning, and predictive/modelling can result in a solid framework for personalized nutrition recommendations. RAG ensures the results retrieved from the previously verified health sources. LoRA/QLoRA achieves the cost-efficient domain adaptation of large language models. LSTM as a sequential model and LightGBM as a structured model accurately forecast calorie intake and nutrient balance. Due to these reasons the above techniques outperform the traditional rule-based nutrition systems with respect to accuracy, personalization, and adaptability

Results show that the response text quality from an integrated framework is good and numerical predictions are credible besides flagging misleading dietary claims as per the BLEU, ROUGE, perplexity, MAE and RMSE study. The generative AIs and predictive analytic jointly work in order to let the system preemptively allow decisions for better ways of living.

In the future, developments will take place that will integrate real-time data from wearables and IoT-empowered nutrition trackers that will add medically validated datasets to the knowledge base and also explainable AI usage. The multi-user technique which supports scalable deployment and the automatic hyperparameter tuning of LoRA/QLoRA adds efficiency and reach for ALICE, whereas reinforcement learning makes the recommendations advance according to user feedback. These improvements will result in a more practical, transparent, and extremely adaptable nutrition recommendation system as a whole.

**VI. REFERENCES**

[1] Addepto Blog Team, “RAG vs. Fine-Tuning: LLM Learning Techniques Comparison,” 2025.

[2] D. M. Anisuzzaman, J. G. Malins, P. A. Friedman, and Z. I. Attia, “Fine-Tuning Large Language Models for Specialized Use Cases,” 2024.

[3] S. Ansari, M. Khan, S. Revankar, A. Varma, and A. Mokhade, “Lightweight Clinical Decision Support System using QLoRA-Fine-Tuned LLMs and Retrieval-Augmented Generation,” arXiv preprint, arXiv:2505.03406, 2025, doi: 10.48550/arXiv.2505.03406.

[4] D. Benfenati, G. M. De Filippis, A. M. Rinaldi, C. Russo, and C. Tommasino, “A Retrieval-Augmented Generation Application for Question-Answering in Nutrigenetics Domain,” Procedia Computer Science, vol. 246, pp. 586–595, 2024, doi: 10.1016/j.procs.2024.09.467.

[5] A. K. Gavai and J. van Hillegersberg, “AI-Driven Personalized Nutrition: RAG-Based Digital Health Solution for Obesity and Type 2 Diabetes,” PLOS Digital Health, vol. 4, no. 5, p. e0000758, May 2025, doi: 10.1371/journal.pdig.0000758.

[6] T. Kayarga, C. Surekha, U. Madeeha, V. G., and V. G., “Nutrivision: Food Calorie Estimation Using Vision Transformers and Personalized Food Recommendations,” in Proc. IEEE Int. Conf. on Data Science, Business, and Systems (ICDSBS), 2025, pp. 1–8, doi: 10.1109/ICDSBS63635.2025.11031737.

[7] S. Khaki, X. Li, J. Guo, L. Zhu, C. Xu, K. N. Plataniotis, A. Yazdanbakhsh, K. Keutzer, and Z. Liu, “SparseLoRA: Accelerating LLM Fine-Tuning with Contextual Sparsity,” arXiv preprint, arXiv:2506.16500, 2025.

[8] D. Khashabi, D. Wadden, N. F. Rajani, Ø. Tafjord, and I. Beltagy, “RAG LLMs Are Not Safer: A Safety Analysis of Retrieval-Augmented Generation LLMs,” 2025.

[9] G. de Castro et al., “Fine-Tuning Large Language Models for Specialized Use Cases,” 2024.

[10] A. Salemi and H. Zamani, “Comparing Retrieval-Augmentation and Parameter-Efficient Fine-Tuning for Privacy-Preserving Personalization of Large Language Models,” 2025.

[11] OpenAI, “A Survey on Parameter-Efficient Tuning of Large Language Models (PEFT),” 2023.

[12] B. Pingua, A. Sahoo, M. Kandpal, D. Murmu, J. Rautaray, R. K. Barik, and M. J. Saikia, “Comparative Analysis of RAG Fine-Tuning and Prompt Engineering in Chatbot Development,” 2024.

[13] B. Pingua, A. Sahoo, M. Kandpal, D. Murmu, J. Rautaray, R. K. Barik, and M. J. Saikia, “Medical LLMs: Fine-Tuning vs. Retrieval-Augmented Generation,” 2025.

[14] S. Pratapa, A. R. Aranha, D. Kumar, G. Malhotra, A. P. N. Iyer, and S. S. Shylaja, “The Fine Art of Fine-Tuning: A Structured Review of Advanced LLM Fine-Tuning Techniques,” 2024.