Fine-Tuning Llama 2 for Automated Ad Caption Generation

Kirtan Panchal[1]; Shubhangi Tidake[2];

[1,2]Professor

[1,2] Vishnupant Potdar Symbiosis Skills and Professional University

Publication Date: 2025/07/15

Abstract: Generating engaging and relevant ad captions poses a significant challenge for advertisers. This research addresses this issue by improving Llama2, an advanced language model, through fine- tuning with a custom dataset created specifically for ad captioning. Techniques such as quantization and matrix decomposition were employed to enhance Llama2's ability to produce captivating and descriptive ad captions.The primary objective was to streamline and improve the efficiency of the caption creation process. Performance evaluation was conducted via A/B testing, comparing our enhanced Llama2 against conventional captioning methods. Key performance indicators included click- through rates, user engagement, and actions taken, such as purchases.Experimental results demonstrated that the fine-tuned Llama2 effectively generates captions that resonate with audiences, encouraging actionable responses. This study advances the capabilities of language models in advertising and provides valuable insights for marketers looking to enhance the impact of their ad campaigns in the digital landscape.

Keywords: *Fine-Tuning, Llama2, Natural Language Processing (NLP), Caption Generation, Advertising, Transfer Learning, Quantization, Text Generation.*

How to Cite: Kirtan Panchal; Shubhangi Tidake; (2025). Fine-Tuning Llama 2 forAutomatedAd Caption Generation. *International Journal of Innovative Science and Research Technology*, 10(7), 740-744. https://doi.org/10.38124/ijisrt/25jul236

I. INTRODUCTION

Large Language Models (LLMs) have recently changed the field of natural language processing (NLP).[1][2], [3] Models like GPT-3 and its successors have large neural networks and are trained on massive datasets, allowing them to understand and create human-like text for many language tasks. Their ability to understand context, find patterns, and give meaningful responses has transformed NLP applications. LLMs have set new standards in language skills and computational linguistics. As technology evolves, the possible uses of LLMs in different areas grow, leading to new solutions for complex language problems and improving AI-driven content creation, analysis, and optimization. Llama2 is a notable example of these advanced models, excelling in language understanding and text generation. This state-of-the-art model is highly useful for various applications, such as text generation and summarization. Its ability to create human-like prose makes it a valuable tool for tasks like writing ad captions that attract and engage target audiences. 4. Creating accurate ad captions is very important in advertising. These short texts grab the attention of potential customers and serve as their first contact with the product or service. Accurate and effective ad captions play a key role in shaping customer opinions and influencing buying decisions. However, making compelling ad captions is challenging. It requires balancing brevity and informativeness, ensuring relevance and appeal, and understanding the target audience, the product, and the platform where the ad will appear. Despite these challenges, automation offers hope for advertisers looking to improve their creative processes. Automated tools and technologies help create ad content more efficiently and consistently, reducing human errors and bias while following brand guidelines. Fine-tuning Llama2 for ad caption generation is a promising approach to using automation to improve the quality and relevance of ad content, increasing audience engagement and achieving better results for advertisers.

II. LITERATURE REVIEW

The study by Simon Lermen, Charlies Rogers-smith, and Jeffrey Ladish (2023) focuses on LoRA fine- tuning and its effects on the Llama 2-Chat 70B model. LoRA is presented as a popular method for fine- tuning large language models. However, the research highlights a crucial concern—this technique can potentially remove the safety features embedded during the original training phase. As a result, the model may become less secure or exhibit undesirable behaviors. The authors emphasize the importance of maintaining safety protocols even while enhancing the model’s performance, offering valuable insights into the balance between model optimization and user safety.

In another 2023 study, T. Dettmers, Artidoro Pagnoni, A. Holtzman, and Luke Zettlemoyer introduce QLoRA, a technique designed for fine-tuning quantized large language models. This method is particularly effective for models that have undergone compression using quantization techniques. By combining LoRA with quantization, QLoRA significantly enhances model efficiency and performance while requiring fewer computational resources. This approach is well-suited for systems that operate with minimal computational resources. Their findings show that QLoRA maintains high accuracy and performance, even under constrained resource conditions, which contributes to its practicality in real- world applications.

A landmark study by Vaswani and colleagues, titled "Attention Is All You Need," (2017) presents the Transformer architecture, which is a cornerstone in modern natural language processing (NLP). The core innovation of this work is the attention mechanism, which allows models to prioritize and focus on relevant parts of the input sequence. Models such as Llama 2 are built upon this framework, which plays a vital role in producing coherent and contextually rich text, including effective captions. Its significance continues to influence a wide range of language model developments.

Lastly, the 2019 research by Dathathri et al. introduces Plug and Play Language Models (PPLM), a technique aimed at controlled text generation. PPLM fine-tunes language models using specific attribute classifiers, enabling the model to generate outputs that align with predefined characteristics without retraining the entire model. This approach supports adaptability and precision in text generation and is useful for tasks like fine-tuning Llama 2 to produce ad captions that meet particular marketing goals.

III. DATA

To fine-tune Llama2, a custom dataset was created by merging data extracted from the company's internal database through an API with additional ad-related content from Facebook's library. This combined dataset featured a wide variety of advertising captions sourced from multiple platforms, resulting in a collection of several hundred examples. It served as a robust foundation for training and evaluating Llama2's capability in generating effective ad captions. Prior to the fine-tuning process, the data underwent thorough preprocessing steps, which included removing irrelevant entries, unifying text formatting, and tokenizing the content to make it compatible with the language model. While the dataset lacked specific labels or annotations, careful attention was given to preserving its consistency and overall quality. Key challenges faced during data preparation included handling varied inputs from multiple origins and ensuring the dataset realistically reflected practical advertising use cases.

IV. METHODOLOGY

- *Fine-Tuning Process for llama2 on the Custom Dataset*

This study adopts a structured approach to fine-tune the Llama 2 model specifically for generating ad captions. The process begins with the collection and cleaning of authentic advertising captions to ensure the dataset is consistent and suitable for training. The fine-tuning phase involves applying efficient techniques such as LoRA and QLoRA, which enhance the model's performance without demanding extensive computational resources. To further optimize the model's efficiency, quantization methods are implemented, enabling it to operate effectively even on low-resource devices. The training process leverages widely used frameworks like Hugging Face and PyTorch, with careful tuning of hyperparameters such as learning rate and batch size to achieve optimal results. After training, the model's effectiveness is assessed using evaluation metrics like BLEU and ROUGE scores, along with real-world indicators such as user engagement and click-through rates. The ultimate aim is to develop a language model that produces engaging, relevant, and brand-appropriate ad captions, all while ensuring responsible and safe language .

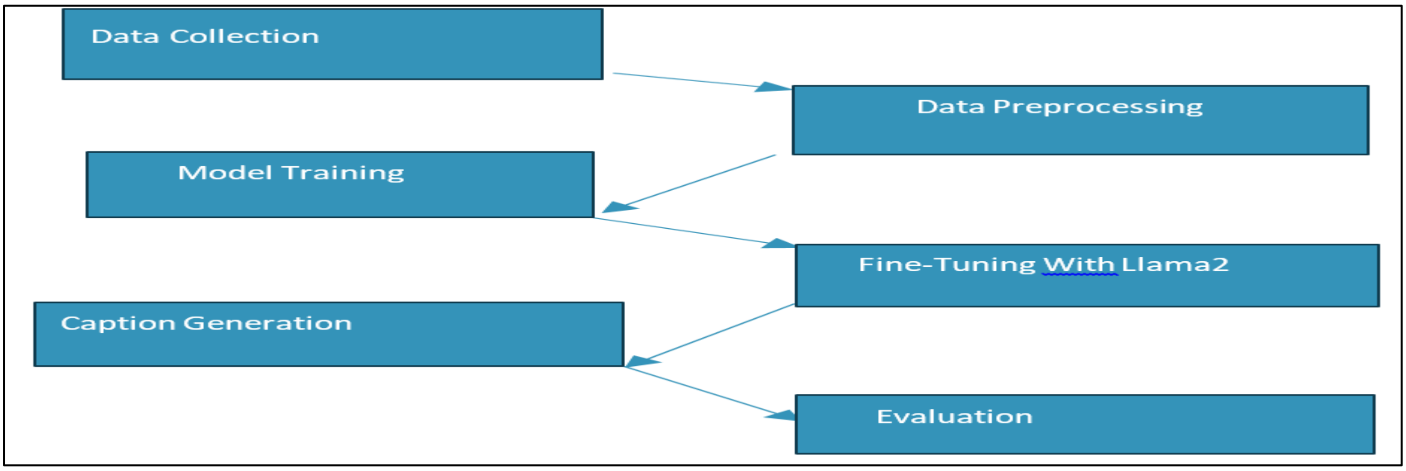


Fig 1 Fine-Tuning Process for llama2 on the Custom Dataset

- ➢ *Fine-Tuning With Lora And Qlora*

Low-Rank Adaptation (LoRA) is a technique that enhances neural network models by adjusting the logits before the softmax layer[15]. This modification aims to improve model performance while reducing memory usage and computational costs. LoRA achieves this by introducing trainable low- rank matrices in specific layers instead of modifying all original weights directly.

- *Basic Structure of Lora:*

- ✓ *Insertion of Low-Rank Matrices:*

LoRA adds trainable low-rank matrices to certain layers of the neural network, which have significantly fewer parameters than the original weight matrices.

- ✓ *Training Process:*

During fine-tuning, only these low-rank matrices are adjusted, while the main weights remain mostly unchanged.

- ✓ *Parameter Efficiency:*

This approach allows for efficient parameter updates, reducing computational resources needed for adjustments.

Quantized Low-Rank Adaptation (QLoRA) builds on LoRA by incorporating quantization, which simplifies numbers without sacrificing model accuracy. QLoRA quantizes the low-rank matrices used in LoRA, reducing the number of bits required to represent these matrices and further decreasing memory usage and computation time[16][17] [18].

- *Basic Structure of Qlora:*

- ✓ *Low-Rank Matrix Insertion:*

Similar to LoRA, QLoRA introduces trainable low-rank matrices into the network's layers.

- ✓ *Quantization Step:*

These low-rank matrices undergo quantization, converting them to a lower precision format to save memory.

- ✓ *Training Process:*

Fine-tuning focuses on adjusting the quantized low-rank matrices while keeping the primary weights largely unchanged.

- ✓ *Enhanced Efficiency*

Quantization enhances efficiency by reducing memory usage and potentially speeding up computations due to faster operations on lower precision numbers.

- ➢ *Description of Hyperparameters*

- *Hyper-parameter Value*

Table 1 Hyperparameters for Fine-tuning

Learning Rate	**2e-4**
Batch Size	4
Epochs	1
Optimizer	paged_adamw_32bit
LoRA Alpha	16
LoRA Dropout	0.1

- ➢ *Evaluation Methodology*

We employed A/B testing to assess the effectiveness of the fine-tuned llama2 model in generating ad captions[18]. This involved comparing the performance of ad captions generated by the fine-tuned model against a baseline model using the original llama2. Specifically, we measured the click-through rate (CTR) of ads with captions from each model[19][20]. By running advertisements with both sets of captions and evaluating their CTRs, we determined which captions were more successful in engaging users and encouraging clicks

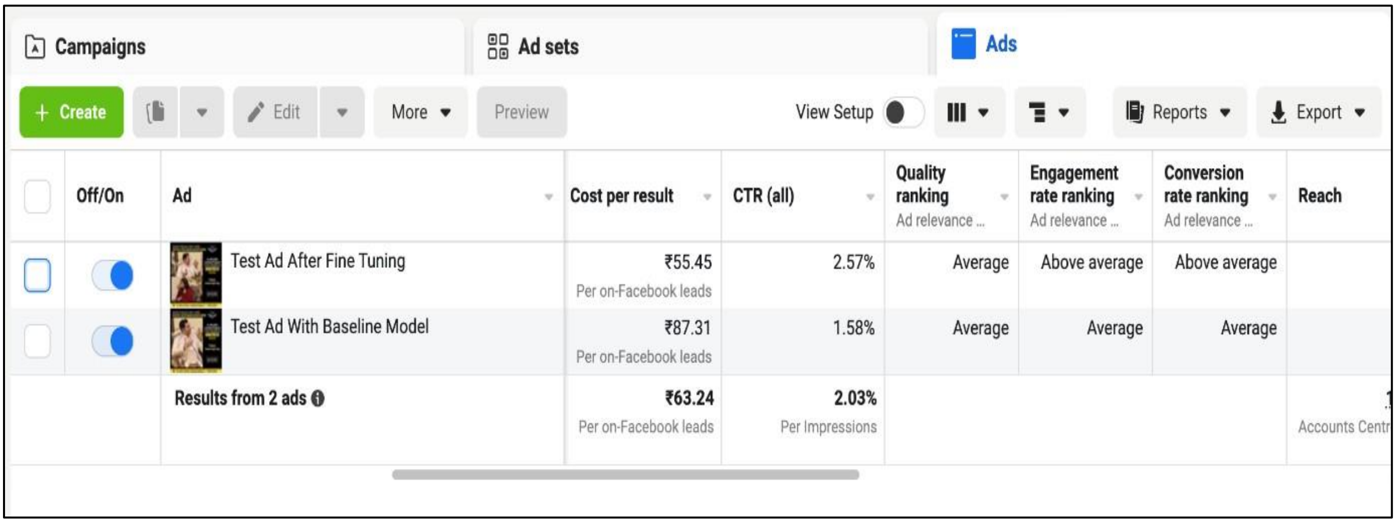
- ➢ *Baseline Model Comparison*

The baseline model used for comparison was the original Lama2 model without specific fine-tuning for ad caption generation. This served as a standard against which we measured the improvements achieved through fine-tuning and evaluated the effectiveness of our approach.

- ➢ *Experimental Setup*

The fine-tuning and evaluation were conducted on an Amazon EC2 server, leveraging its computational capabilities to facilitate efficient model training and testing. This setup provided the necessary infrastructure to handle the computational demands of fine-tuning a large language model like llama2, ensuring scalability and reliability throughout the experiment.

V. RESULT



□	Off/On	Ad	Cost per result	CTR (all)	Quality ranking Ad relevance ...	Engagement rate ranking Ad relevance ...	Conversion rate ranking Ad relevance ...	Reach
□	●	Test Ad After Fine Tuning	₹55.45 Per on-Facebook leads	2.57%	Average	Above average	Above average	
□	●	Test Ad With Baseline Model	₹87.31 Per on-Facebook leads	1.58%	Average	Average	Average	
		Results from 2 ads	₹63.24 Per on-Facebook leads	2.03% Per Impressions				Accounts Centr

Fig 2 Ad Caption Generation without Fine-Tuning Llama2 Model

- Run Ad With Llama 2 Baseline Model:-

CTR : 1.58%

Cost per result(Lead) : 87.31 rs.

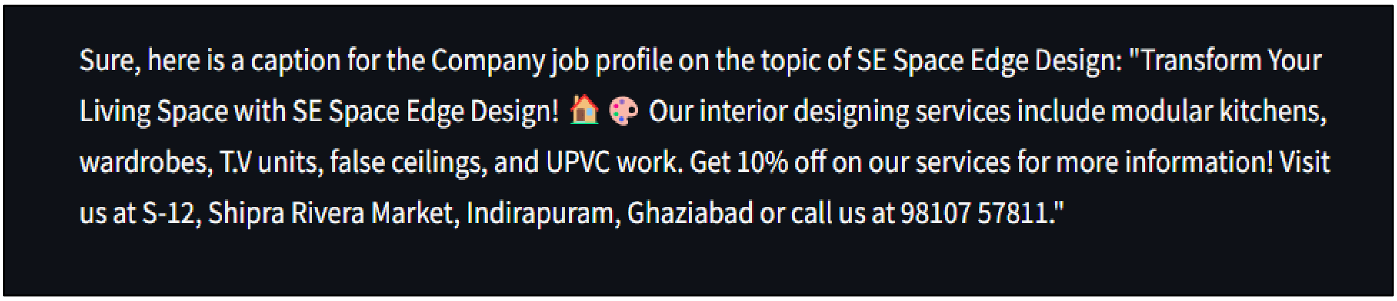
Customer Engagement rate : Average

- Run Ad With Llama 2 Fine-Tuned Model:-

CTR : 2.57%

Cost per result(Lead) : 55.45 rs.

Customer Engagement rate : Above Average



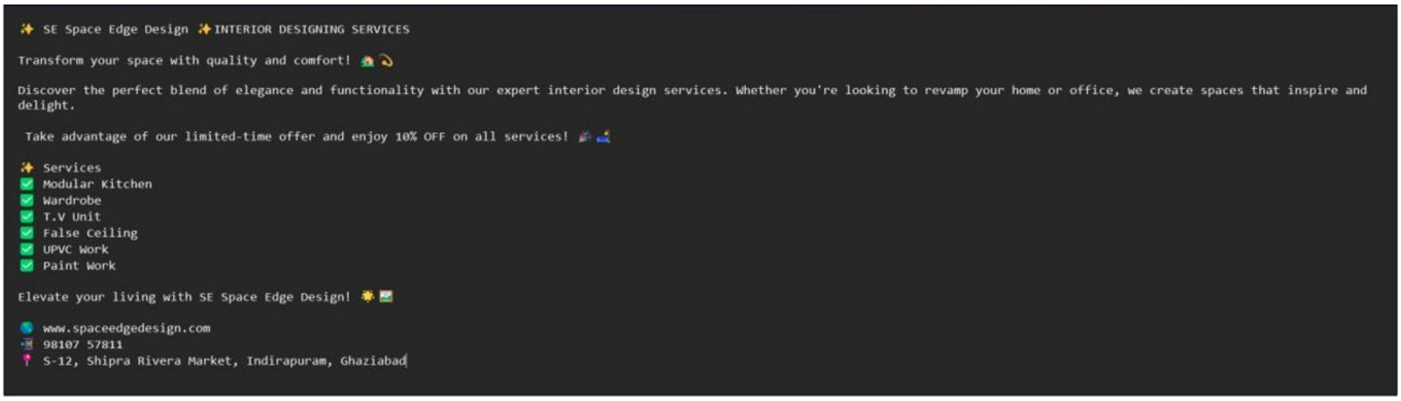
Sure, here is a caption for the Company job profile on the topic of SE Space Edge Design: "Transform Your Living Space with SE Space Edge Design! 🏠 🎨 Our interior designing services include modular kitchens, wardrobes, T.V units, false ceilings, and UPVC work. Get 10% off on our services for more information! Visit us at S-12, Shipra Rivera Market, Indirapuram, Ghaziabad or call us at 98107 57811."

Fig 3 Run Ad with Llama 2 Baseline Model

This figure displays the results of generating ad captions using the Llama2 model before any fine- tuning. The captions produced at this stage are generally generic and lack the specific details needed for effective advertising campaigns.

✓ *Purpose*

This figure aims to establish the baseline performance of the Llama2 model without fine- tuning. A comparison with the results in Figure 2 highlights the improvements achieved through the fine-tuning process.



```
✨ SE Space Edge Design ✨INTERIOR DESIGNING SERVICES

Transform your space with quality and comfort! 🏡💫

Discover the perfect blend of elegance and functionality with our expert interior design services. Whether you're looking to revamp your home or office, we create spaces that inspire and
delight.

 Take advantage of our limited-time offer and enjoy 10% OFF on all services! 🎉🛋

✨ Services
✅ Modular Kitchen
✅ Wardrobe
✅ T.V Unit
✅ False Ceiling
✅ UPVC Work
✅ Paint Work

Elevate your living with SE Space Edge Design! 🌟🖼

🌐 www.spaceedgedesign.com
📲 98107 57811
📍 S-12, Shipra Rivera Market, Indirapuram, Ghaziabad
```

Fig 2 Ad Caption Generation with Fine-Tuned Llama2 Model

This figure illustrates the outcomes of generating ad captions after fine-tuning the Llama2 model with a specialized dataset. The captions generated in this phase are more relevant, context-specific, and effective for targeted advertising.

✓ *Purpose:*

The inclusion of this figure demonstrates the effectiveness of the fine-tuning process. Comparing Figures 1 and 2 provides conclusive evidence of how fine-tuning enhances the model's capability to generate high-quality and pertinent ad captions.

VI. LIMITATION AND FUTURE WORK

Despite achieving positive results, our research faced several limitations. These included constraints related to data availability, model complexity, and computational resources. Overcoming these challenges presents opportunities for future studies to enhance fine-tuning techniques, incorporate multimodal inputs, and explore larger datasets to improve the accuracy of ad caption generation.

VII. CONCLUSION

This research examined the effectiveness of fine-tuning the llama2 language model for generating ad captions and compared its performance with baseline models.[21][22][23] Our findings demonstrate that fine-tuning llama2 significantly enhances ad caption generation by producing more accurate and relevant captions compared to standard models. These results underscore the model's potential to improve advertising communication, leading to increased audience engagement and click-through rates. Moreover, our study highlights the importance of expanding natural language processing (NLP) applications in advertising. By adapting pre-trained models like llama2 through fine-tuning, advertisers can create more effective and appealing ad campaigns. This research paves the way for future innovations in NLP for advertising, showcasing the transformative potential of these technologies in shaping marketing strategies.

REFERENCE

[1]. H. Touvron et al., “Llama 2: Open foundation and fine-tuned chat models,” arXiv preprintarXiv: 2307.09288, 2023.

[2]. S. Sehgal, J. Sharma, and N. Chaudhary, “Generating image captions based on deep learning and natural language processing,” in 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO), 2020, pp. 165–169.

[3]. J. Wei et al., “Emergent abilities of large language models,” arXiv preprintarXiv: 2206.07682, 2022.

[4]. W. Yu et al., “A survey of knowledge-enhanced text generation,” ACM Comput Surv, vol. 54, no. 11s, pp. 1–38, 2022.

[5]. M. H. Bakri, “The effectiveness of advertising in digital marketing towards customer satisfaction,” Journal of Technology Management and Technopreneurship (JTMT), vol. 8, no. 1, pp. 72–82, 2020.

[6]. C. Jeong, “Fine-tuning and utilization methods of domain-specific llms,” arXiv preprintarXiv: 2401.02981, 2024.

[7]. S. Lermen, C. Rogers-Smith, and J. Ladish, “Lora fine-tuning efficiently undoes safety training in llama 2-chat 70b,” arXiv preprint arXiv:2310.20624, 2023.

[8]. T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, “Qlora: Efficient finetuning of quantized llms,” Adv Neural Inf Process Syst, vol. 36, 2024.

[9]. V. Ashish, “Attention is all you need,” Adv Neural Inf Process Syst, vol. 30, p. I, 2017.

[10]. S. Dathathri et al., “Plug and play language models: A simple approach to controlled text generation,” arXiv preprint arXiv:1912.02164, 2019.