Comparative Analysis of Finetuned Language Models for Ramayana

Jyotiraditya J
pes1202100742@pesu.pes.edu

Sandeep Ram
pes1202101992@pesu.pes.edu

K Ganesh Vaidyanathan
pes1202100629@pesu.pes.edu

Preethi P
preethip@pesu.pes.edu

April 24, 2024

Abstract

In this paper, we present a comparative study of various fine tuned Language Models (LLMs) for educational applications specifically related to answering questions on the great Indian epic Ramayana. The goal of this comparative study is to analyse the ability of different LLMs in identifying key pieces of information from a given textual corpus as well as deriving reason and attempting to uncover deeper meanings from the core text. We fine tuned Gemma 2b, NurtureAI, Openchat, Orca, Vicuna, Llama, and Zephyr using the QLora fine tuning technique, leveraging various interpretations of the Ramayana text available online as our datasets. We assess their performance using deepeval, analyzing metrics such as hallucination, faithfulness, contextual relevancy, answer relevancy, and toxicity. Our findings provide insights into the effectiveness of these models in generating coherent and informative responses to questions related to the Ramayana, offering valuable guidance for their deployment in educational settings.

Keywords: *finetuned language models, Ramayana, educational applications, deepeval, QLora, comparative analysis*

1 Introduction

The rapid advancements in natural language processing have led to the development of powerful Language Models (LMs) capable of understanding and generating human-like text. These LMs hold significant potential for enhancing various applications in education, including question answering, essay grading, and personalized tutoring. In this paper, we focus on evaluating the performance of several finetuned LLMs for educational applications specifically related to answering questions on the great Indian epic Ramayana. We utilized various interpretations of the Ramayana text available online as our datasets for finetuning the models.

2 Related Work

Previous research has investigated the use of LLMs in educational settings, highlighting their potential to automate grading, provide personalized feedback, and facilitate interactive learning experiences. However, limited attention has been given to comparing the performance of different LLMs in educational applications, particularly concerning their ability to generate contextually relevant and accurate responses to questions related to specific literary texts such as the Ramayana. Our study builds upon existing literature by conducting a comprehensive comparative analysis of finetuned LLMs, leveraging state-of-the-art evaluation techniques to assess their effectiveness in the context of the Ramayana.

2.1 Gemma Models

: we use Gemma2B, a family of open models based on Google's Gemini models, with strong performance in language understanding, reasoning, and safety. Gemma outperforms similarly sized models on various text-based tasks and includes comprehensive safety evaluations

2.2 Orca 2B

: Orca 2 is designed to excel in reasoning tasks. It provides single-turn responses in various tasks, including reasoning over user-provided data, reading comprehension, math problem solving, and text summarization.Orca 2B is the finetuned version of the LLAMA 2 model.

2.3 Mistral 7B

: Mistral 7B outperforms the best open 13B model and the best released 34B model in various benchmarks, especially in reasoning, mathematics, and code generation. he model is fine-tuned to follow instructions, resulting in the Mistral 7B–Instruct, which surpasses other models in both human and automated benchmarks.

2.4 Llama 2 7B

: Llama 2 7B utilizes an optimized transformer architecture designed for auto-regressive language generation. This model has been fine-tuned using techniques like SFT (Supervised Fine-Tuning) and RLHF (Reinforcement Learning from Human Feedback), enhancing its alignment with human-like text generation in terms of helpfulness and safety.

2.5 Zephyr 7B

: ZEPHYR-7B is a powerful 7B parameter language model designed for user intent alignment. Unlike other models, it achieves impressive performance without relying on human annotation. The secret lies in its innovative approach: distilled direct preference optimization (dDPO). ZEPHYR-7B outperforms larger models on various benchmarks, making it a valuable tool for natural language understanding and generation tasks

3 Problem Statement Formulation

The primary objective of this study is to compare the performance of various finetuned LLMs in generating responses to questions related to the great Indian epic Ramayana. Specifically, we aim to evaluate their ability to produce coherent, contextually relevant, and accurate answers while minimizing hallucination and toxicity. By assessing these aspects, we seek to identify the strengths and weaknesses of each model, providing insights into their suitability for different educational tasks and domains.

4 Methodology

For this study, we employed the QLora finetuning technique to adapt the aforementioned language models to the domain of the Ramayana, a renowned Indian epic. We utilized a diverse range of interpretations and datasets related to the Ramayana text, available

online, as our finetuning corpus. The finetuning process aimed to optimize these models' performance in answering queries specifically related to the Ramayana epic, leveraging the RAG pipeline implementation. We have used this technique to fine-tune Gemma 2b, NurtureAI, Openchat, Orca, Vicuna, Llama and Zephyr LLM's, each of which are around 7 billion parameters, allowing us to compare the performance of similarly sized LLM's.

4.1 QLora Finetuning

QLora (Quantized Lotra) is a finetuning technique that allows for efficient adaptation of large language models to specific domains or tasks. It involves quantizing the finetuned weights, reducing their precision from the standard 32-bit floating-point format to lower bit-widths (e.g., 8-bit or 4-bit). This quantization process significantly reduces the memory footprint and computational overhead, making it feasible to finetune and deploy large models on resource-constrained devices or environments.

4.2 Retrieval-Augmented Generation (RAG)

The Retrieval-Augmented Generation (RAG) pipeline is a technique that combines the strengths of retrieval-based and generation-based approaches for language tasks. It consists of two main components: a retriever and a generator. The retriever component searches through a large corpus of text and retrieves relevant passages or documents based on the input query. These retrieved passages are then passed to the generator component, which is a large language model trained to generate coherent and contextual responses based on the input query and the retrieved passages.

5 Experiments and Results

We conducted extensive experiments to assess the performance of the finetuned LLMs using deepeval specifically in the context of answering questions on the Ramayana. Our results indicate significant variations in the models' performance across different metrics. For example, Gemma 2b exhibited high contextual relevancy and answer relevancy scores but struggled with hallucination and toxicity when answering queries related to the Ramayana. In contrast, NurtureAI demonstrated superior faithfulness and low toxicity but lagged in contextual relevancy in the context of the Ramayana. Openchat performed well in generating diverse responses but showed tendencies towards hallucination when answering questions on the Ramayana. Orca, Vicuna, Llama, and Zephyr displayed mixed performance across the evaluated metrics, highlighting the need for nuanced considerations when selecting LLMs for specific educational tasks related to the Ramayana.

5.1 Questions

some of the questions we used are like this:

1. Who is Hanuman?
2. What are the principles of Lord Ram one can draw inspiration from?
3. What was the name of the forest where Lord Rama, Lakshmana and Goddess Sita stayed during exile?

4. Discuss the concept of dharma (duty/righteousness) as it is portrayed in the Ramayana. How do characters navigate conflicting duties, and what are the consequences of their choices?

5. How does the portrayal of masculinity and femininity in the Ramayana reflect societal norms and expectations during the time of its composition?

5.2 Metrics

In this study, we evaluate the performance of language models using four key metrics: Answer Relevancy, Faithfulness, Contextual Relevancy, and Toxicity. These metrics are defined as follows:

Answer Relevancy measures the quality of your RAG pipeline's generator by evaluating how relevant the actual output of your LLM application is compared to the provided input.

Faithfulness measures the quality of your RAG pipeline's generator by evaluating whether the actual output factually aligns with the contents of your retrieval context.

Contextual Relevancy measures the quality of your RAG pipeline's retriever by evaluating the overall relevance of the information presented in your retrieval context for a given input.

Toxicity quantifies the presence of harmful, offensive, or inappropriate content in the model's response. A lower toxicity score is desirable, indicating that the model's output is less likely to contain toxic or undesirable language.

Table 1: Performance of each LLM on the Factual Questions

Model	Toxicity	Hallucination	Contextual Relevancy	Faithfulness	Answer Relevancy
Gemma 2B	0.00	1.50	1.00	0.77	0.80
NurtureAI	0.00	1.66	1.50	1.30	1.66
Openchat	0.00	1.66	1.00	1.231	1.333
Orca	0.00	1.50	1.50	1.50	1.50
Vicuna	0.00	1.66	1.50	1.50	1.66
Llama	0.00	2.00	0.50	1.40	2.00
Zephyr	0.00	1.00	0.50	1.875	1.20

Table 2: Performance of each LLM on the Inference questions

Model	Toxicity	Hallucination	Contextual Relevancy	Faithfulness	Answer Relevancy
Gemma 2B	0.00	0.00	1.00	0.75	0.00
NurtureAI	0.00	1.00	0.50	1.00	0.833
Openchat	0.00	1.00	0.50	0.714	1.00
Orca	0.00	1.00	0.50	0.50	1.00
Vicuna	0.00	1.00	1.00	1.00	1.00
Llama	0.00	1.00	1.00	1.00	0.80
Zephyr	0.00	0.50	0.50	0.70	1.00

Table 3: Performance of each LLM on the Long Answer questions

Model	Toxicity	Hallucination	Contextual Relevancy	Faithfulness	Answer Relevancy
Gemma 2B	0.00	2.00	1.00	2.00	0.00
NurtureAI	0.00	1.00	1.50	1.833	2.00
Openchat	0.00	2.00	1.00	2.00	2.00
Orca	0.00	1.00	1.00	1.928	2.00
Vicuna	0.00	0.50	1.50	2.00	2.00
Llama	0.00	2.00	1.00	2.00	1.875
Zephyr	0.00	2.00	1.50	2.00	1.888

6 Conclusions

In conclusion, our comparative analysis of finetuned LLMs for educational applications related to the Ramayana provides valuable insights into their strengths and limitations. While each model exhibits unique capabilities, none emerges as universally superior across all evaluated metrics. Educators and practitioners must carefully evaluate the trade-offs between coherence, relevancy, faithfulness, and toxicity when selecting LLMs for educational tasks related to the Ramayana. Future research directions include exploring ensemble approaches and domain-specific finetuning strategies to further improve the performance of LLMs in educational contexts.

References

Gustavo Aguilar, Suraj Maharjan, Adrian Pastor López Monroy, and Thamar Solorio. A Multi-task Approach for Named Entity Recognition in Social Media Data. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, Association for Computational Linguistics, 2017.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. Multimodal Named Entity Recognition for Short Social Media Posts. 2018. https://arxiv.org/abs/1802.07862.

Ron Mokady, Amir Hertz, and Amit H. Bermano. ClipCap: CLIP Prefix for Image Captioning. 2021. https://arxiv.org/abs/2111.09734.

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. CRIS: CLIP-Driven Referring Image Segmentation. 2022. https://arxiv.org/abs/2111.15174.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. Zephyr: Direct Distillation of LM Alignment. *arXiv preprint arXiv:2310.16944*, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross

Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*, 2023.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi, Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah. Orca 2: Teaching Small Language Models How to Reason. *arXiv preprint arXiv:2311.11045*, 2023.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu-hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma: Open Models Based on Gemini Research and Technology. *arXiv preprint arXiv:2403.08295*, 2024.