5th International Conference on Advanced Research in Computing (ICARC) 2025
Graduate Colloquium

Paper ID:80

Towards Ethical Inference in Language Models: Integrating Religious Data and Enhancing Responsible LLM Development

Ranasinghe K.S.[1*], Paik I.[1]

[1]*University of Aizu, Japan*
[*]krsranasinghe@gmail.com

Large Language Models (LLM) have emerged as the most powerful tools to perform various aspects in daily life. These models are capable of diverse tasks including text understanding and generating, image generation, language translation and sentiment analysis. Continual advancements in LLM are expanding the scope of their capabilities enabling wide range of applications. Although LLMs have made significant progress, still there are challenges and limitations that needs to be addressed. As the existing LLM models generally focus on the natural language processing related tasks, it is crucial to emphasize the training and fine-tuning of ethical LLMs. When developing and fine-tuning LLMs, issues such as biased responses and lack of moral consistency can arise. This could lead to significant ethical challenges, particularly because the data used for training heavily influences the model's outputs. Developing a specific ethical LLM by establishing a benchmark for ethical performance could help overcome this problem. The primary goal of this research is to implement an ethical inferences language model which can make the predictions based on the religious data. Religious data is used for the fine-tuning and Llama-2-7B-chat model is used along with Low Rank Adaptation techniques. The fine-tuned model was tested by generating the responses to prompts related to ethical scenarios and the accuracy of the model can be calculated. The model trained with 5000 Bible data. During the training loss decrease gradually by denoting the model learns well with the data. The fine-tuned model provides reliable performance when working with ethics-related data. Further the Fine-tuned model demonstrated the ability to generate text based on ethical prompts, showing a positive trend in the generated ethical inferences indicating that this model can be developed further by training with more religious data from Bible, Quran, Hindu scriptures and Tripitaka. In future the model will be refined further using Supervised Fine Tuning to obtain more accurate model with enhanced ethical inference capabilities.

Keywords: *Fine-tuning, Large Language Model, Llama 2, Religious Data*