Project Report

1. Introduction

The advent of Large Language Models (LLMs) has marked a transformative era in artificial intelligence, enabling machines to perform complex language tasks with unprecedented efficiency and nuance. LLMs such as Llama-2 have demonstrated a remarkable ability to generate text that closely mimics human writing, understand context, and provide answers that are both informative and contextually relevant. Their application spans various fields, from creative writing to technical support, and more recently, into the realm of policy-making and analysis.

Recognizing the potential of LLMs to revolutionize the field of AI policy, our team has embarked on an ambitious project: the creation of an AI Policy Chatbot. This chatbot, dubbed "AI Policy Chat," aims to serve as a specialized assistant capable of dissecting, interpreting, and elucidating AI policy-related inquiries. By leveraging the advanced text processing capabilities of LLMs, the AI Policy Chat will offer in-depth insights, guiding users through the intricacies of AI regulations, ethical considerations, and governance frameworks.

The motivation behind this endeavor stems from a pressing need for tools that can demystify the often complex language of AI policy, making it accessible to a broader audience. There exists a gap in the industry for such a dedicated platform, and by harnessing our collective knowledge of LLMs, we aspire to bridge this gap.

2. Methodology

Our methodology for developing the AI Policy Chat leveraged comprehensive data collection and meticulous model fine-tuning. Initially, we aggregated over 100 policy reports through web scraping and meticulously crafted nearly 500 Q&A pairs to form our training dataset. This rich corpus served as the foundation for training our model to understand and respond to queries about AI policy with precision.

In the model development phase, we employed the Llama-2 7b model, selected for its balance of size and performance. Fine-tuning involved tweaking various hyperparameters, including epochs, to refine the model's responses. The optimized version of our model, which showed superior performance, has been shared on the Hugging Face platform, making it accessible for public use and further research.

Our prompting techniques were diverse, encompassing chain of thoughts, tree of thoughts, priming, examples, and task instruction strategies. These techniques were instrumental in guiding the model to generate well-structured and contextually relevant outputs.

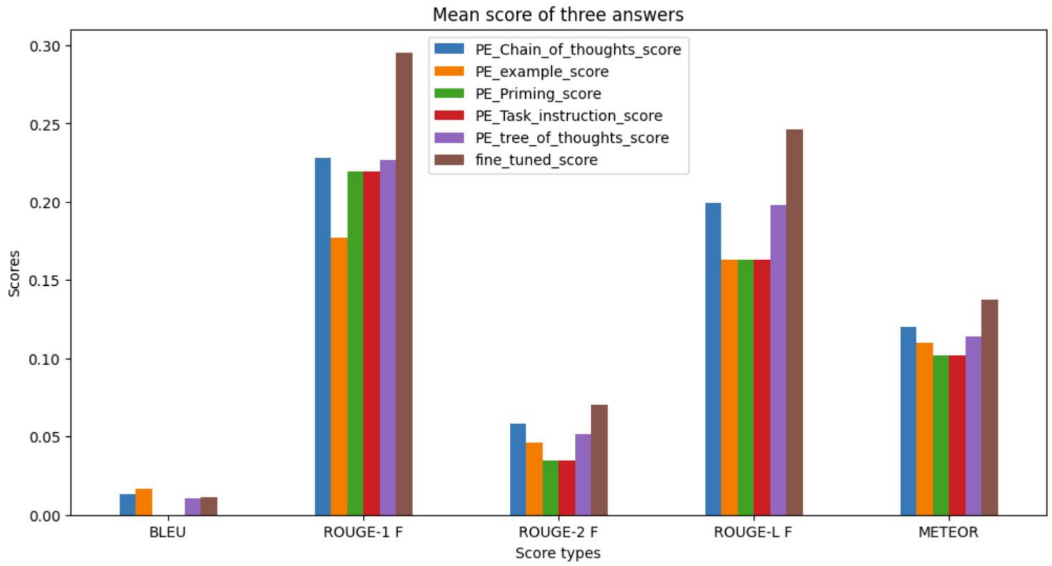
For evaluation, we employed a mix of automated metrics such as BLEU, ROUGE, and METEOR scores to objectively assess the linguistic quality of the generated text.

Complementing these were human evaluations focused on relevance, correctness, and informativeness(each of manual score is in the range of 0 to 10, 10 means that the answer has very strong relevance, correctness, and informativeness), ensuring our model's outputs align with the nuanced demands of AI policy discourse. Through this rigorous process, we aimed to create a tool that not only understands the complexities of AI policy but also communicates them effectively.

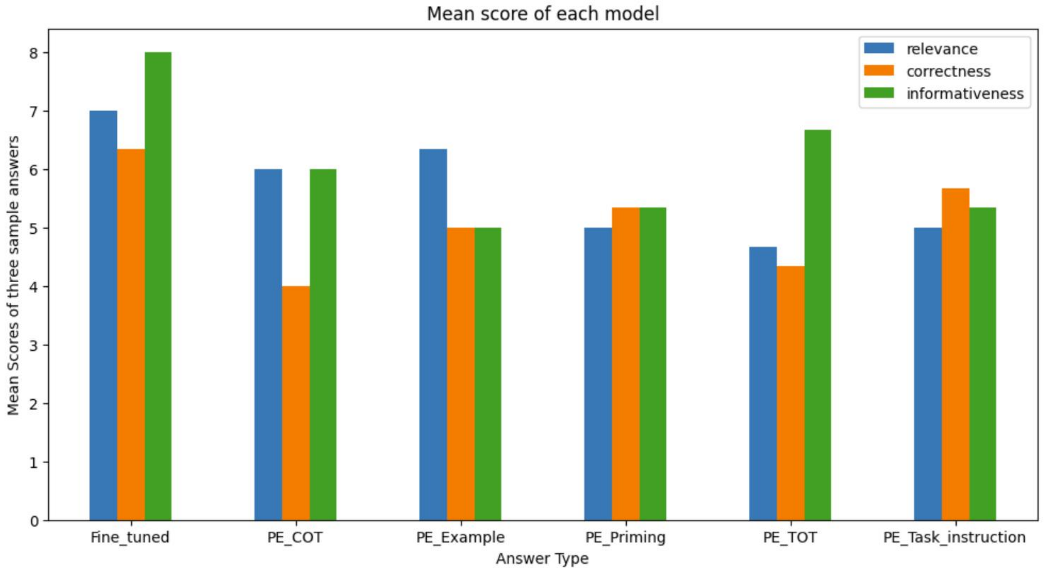
3. Results

We evaluate our models through mean score of answers generated for three random questions for each model adjustment methods(five prompt engineering and one fine tuning). And our results are as follows:

Automatic scores for each methods:



Human evaluation score for each methods:



After carefully comparing these scores, we found that for the automatic scores, apart from the BLEU, the fine-tuned score is always the highest among all the adjustment methods. And second highest is Chain of thoughts prompt engineering.
The BLEU scores for each methods of the adjustment are very low, it might due to the limitations of score itself.

For the human evaluation scores, fine-tuned model generated answers are higher than other methods among the relevance, correctness, and informativeness. Tree of thoughts prompt engineering performs well in informativeness. Example and chain of thoughts prompt engineering performs well in relevance. Task instruction prompt engineering performs well in correctness.

Key findings and insights: We would get answers of better relevance, correctness, and informativeness score and automatic scores after fine-tuning the model. However, prompt engineering is also important in the model training part since it gives us the answer pattern we want. The model could be improved greatly after prompt engineering and fine tuning of the parameters.

4. Discussion

Strengths and weaknesses of different approaches: Fine tuning could be more accurate, relevant and correct, but it usually takes more time and we need to adjust many parameters to try to find the optimal combinations. Prompt engineering can help us generate the answer in the pattern we want, but it lacks high score of evaluations. If we could include RAG to add more specific policy domain knowledge, the relevance, informativeness would be increased greatly.

Challenges encountered and solutions:
1. We were faced with limited available GPU, So we split the task into different part by using Google Colab in different account.
2. It is hard to connect the llama2 weights to the colab, so we used Vanilla model from hugging face rather than downloading the big weights from the website.
3. Some Prompt engineering and fine-tuning process is time consuming, so we do inconflict tasks at the same time.
4. Automatic scores are lack of accuracy, and human evaluation scores are very diverse and personal to ensure the integrity of the score.

Implications for AI policy development and research: It could help the policy makes subtract relevant data and policy quickly without searching online for a long time.

5. Conclusion

The project report details the development and evaluation of an AI Policy Chatbot, aimed at simplifying AI policy discussions for a wider audience by leveraging Large Language Models

(LLMs), specifically using Llama-2. This initiative addresses the need for accessible tools in understanding AI policy, thereby filling an existing gap in the industry.

Summary of Key Takeaways

Development Approach: The project utilized a comprehensive methodology involving extensive data collection, including over 100 policy reports and nearly 500 Q&A pairs. The Llama-2 7b model was fine-tuned with a focus on diverse prompting techniques and hyperparameter optimization, enhancing its ability to deliver precise and contextually relevant responses on AI policy.

Evaluation Results: Evaluation encompassed both automated metrics (BLEU, ROUGE, METEOR) and human assessments on relevance, correctness, and informativeness. Fine-tuning was identified as significantly enhancing the model's performance across these metrics, compared to various prompt engineering methods. Among prompt engineering strategies, Chain of Thoughts and Tree of Thoughts showed particular strengths in certain areas (e.g., relevance, informativeness).

Strengths and Limitations: The report highlights fine-tuning for its accuracy and relevance, despite the time and resource intensity required. Prompt engineering offers quicker, pattern-specific responses but lacks in evaluation scores. Challenges such as limited GPU resources, model integration, and the time-consuming nature of fine-tuning and prompt engineering were addressed through practical solutions.

Implications and Recommendations: The AI Policy Chat demonstrates potential to assist policy analysis by quickly providing relevant, correct, and informative answers, suggesting its utility for policymakers. Future work should explore incorporating Retrieval-Augmented Generation (RAG) for domain-specific knowledge enhancement and further optimize prompt engineering and fine-tuning techniques to improve performance.

Recommendations for Future Work

Incorporate Domain-Specific Knowledge: Implementing RAG or similar techniques could significantly boost the relevance and informativeness of responses by integrating more specific AI policy knowledge.

Optimization of Prompt Engineering and Fine-Tuning: Continued experimentation with prompt engineering and fine-tuning parameters can help in identifying optimal strategies for improving model responses.

Expand Training Data: Enriching the training dataset with more diverse policy documents and Q&A pairs can enhance the model's understanding and ability to generate nuanced responses.

Enhance Evaluation Methods: Developing more sophisticated evaluation metrics or methodologies could provide deeper insights into the model's performance and areas for improvement.

Explore Scalability Solutions: Addressing resource limitations, such as GPU availability, through scalable solutions or partnerships can facilitate more extensive model training and refinement.

By addressing these recommendations, future iterations of the AI Policy Chat can achieve greater efficacy and impact, making AI policy discussions even more accessible and informative.