AI Policy Chat

94812-A3

Applications of NL(X) and LLM – Final Project

Spring 2024

Ashutosh Choudhari, Dibyanshu Patnaik, Gautam Devadiga,
Santhoshkumar Panneerselvam, Adam Patula,
Chitra Raghavendrarao Krishnarao

Instructed by Professor Anand Rao

Table of Contents

1. Introduction

1.1 Project Motivation and Objectives

With the rapid advancement in AI and the development of new tools, it's essential to develop a guiding framework concerning its implementation and usage. This framework should address the potential risks in terms of AI security, ethical, legal, and data protection, yet it should not be so restrictive that it suppresses innovation. Since AI systems require an abundance of data for training, adhering to data protection regulations is a priority for the company, especially while handling sensitive information. A potential data breach could severely impact the organization leveraging the AI tool; it could not only result in legal issues but also loss of trust and customers. However, placing severe restrictions is not a viable option as AI is starting to be built into existing tools. Therefore, having an AI policy is a necessity; it will act as a roadmap for organizations to leverage the power of AI while upholding ethical standards, complying with regulations, and fostering trust among stakeholders

The objective of this project is to develop a fine-tuned Large Language Model titled 'AI Policy Chat' that can comprehend, and address queries related to AI Policy. We have made use of LLMs for text generation and to ensure that the response provided is contextually relevant and grammatically accurate. We will explore three architectures to accomplish this:

(i) Prompt engineering with a base Llama-2 7b model

(ii) Retrieval Augmented Generation

(iii) Fine-tuned base Llama-2 7b model

To achieve our objectives, the project will be divided into two main phases:

1. Data collection and model development: Gather relevant datasets and design the initial model architectures.

2. Model fine-tuning and performance evaluation: Refine the models and rigorously assess their performance against predefined metrics.

1.2 Background on AI Policy and LLMs

1.2.1 AI Policy

A good AI policy will establish guidelines for data collection, storage, and usage, ensuring compliance with data protection regulations and safeguarding sensitive information. In October 2018, over 250 experts and 60 organizations, representing more than 40 countries, endorsed the Universal Guidelines for Artificial Intelligence ("UGAI"). The guidelines were organized by the Public Voice. The guidelines comprise the following areas: Right to Transparency, Right to Human Determination, Identification Obligation, Fairness Obligation, Assessment and Accountability Obligation, Accuracy, Reliability, & validity Obligations, Data Quality Obligation, Public Safety Obligation, Cybersecurity Obligation, Prohibition on Secret Profiling, Prohibition on Unitary Scoring, and Termination Obligation.

The OECD(Organization of Economic Cooperation and Development AI Principles) AI Principles were adopted in 2019 and endorsed by 42 countries—including the United States, several European Countries, and the G20 nations. The guidelines comprise the following areas: Inclusive growth, sustainable development and well-being, Human-centered values and fairness, Transparency and explainability, Robustness, security and safety, and Accountability.

1.2.2 Large Language Models

Large language models (LLMs) are very large deep-learning models that are pre-trained on extensive datasets to achieve high flexibility. These models excel at a variety of text-generation tasks such as question-answering, summarization, translation, and text-completion. Models such as ChatGPT, Claude 2, and Llama 2 that can efficiently indulge in question-answering and text-generation hint that while they're inching closer towards human-like performance, they could potentially outperform them.

Therefore, LLMs can be leveraged for AI policy analysis and decision-making. Since LLMs can understand complex legal texts and identify trends across multiple documents, policymakers, and stakeholders can leverage them for AI policy analysis and decision-making. By doing this, not only will significant time be saved, but also any potential discrepancies in understanding the AI policy documents will be minimized.

2. Methodology

2.1 Data Collection and Preparation

The data collection process involved gathering approximately 4 AI policy documents from each of the appendices titled B1-B5, totaling 20 documents overall. Subsequently, 100 question-answer (<Q, A>) pairs were manually created overall by examining the AI policy documents. These pairs were combined with those formulated by other teams, resulting in a total of 436 pairs. The pairs were presented in both JSONL and JSON formats, the latter was used for model development and evaluation. A 60-20-20 split was applied to allocate the (<Q, A>) pairs for training, validation, and testing, respectively.

2.2 Fine-Tuning and Model Development

Four distinct approaches were employed for model development and fine-tuning. Firstly, the Llama-2 7b model was utilized; this takes the question-answer (<Q, A>) pairs as input and generates responses based on similarity to the text in the AI Policy documents. Secondly, this was fine-tuned to enhance its performance for this particular task.

Subsequently, the Retrieval Augmented Generation (RAG) model was employed; this uses the (<Q, A>) pairs to derive context. This is then combined with the original input and fed to the LLM in order to generate responses. Furthermore, this model too was fine-tuned to improve its performance.

These approaches offer a diverse range of strategies for fine-tuning and enhancing the base LLM's capabilities in generating contextually relevant responses to AI policy-related queries. Each model variation was evaluated based on predefined metrics to assess its effectiveness in meeting our project objectives.

2.3 Prompting Techniques

The methods included for prompt engineering include zero-shot, one-shot, few-shot, negative prompting, and chain-of-thought prompting.

Zero-shot prompting involved providing no additional context or prompts to the model during inference, allowing it to generate responses solely based on the input query. For example, Input: List the most pressing topics regarding regulation of AI.

Example Output -

Q: How might AI affect competition between nations in the future?

Base Model:

A: How might AI affect competition between nations in the future? What are the implications for the future of warfare? How might AI affect competition between nations in the future? What are the implications for the future of warfare? How might AI affect competition between nations in the future? What are the implications for the future of warfare?

Fine-Tuned Model:

A: AI is a general-purpose technology that can be used across many industries and applications. It can be used to increase productivity, reduce costs, and enable entirely new products and services. As such, it can be a powerful tool for competitive advantage in the global economy. In the future, AI could be used by nations to gain an edge in areas like defense, where AI can help improve the accuracy and speed of targeting and reduce the costs of deploying miss

One-shot and few-shot prompting involved supplying the model with a small number of context or prompt examples to guide its response generation process.

(One-shot) For example, Task: You are a chatbot named AI PolicyChat. Provide appropriate commentary on discussions about legal, ethical, and moral issues regarding the development, testing, evaluation, and use of Generative AI:

Example: What is the purpose of the Executive Order on the safe, secure, and trustworthy development and use of artificial intelligence?

The purpose of the Executive Order is to guide the development and use of artificial intelligence (AI) in a manner that is safe, secure, and trustworthy. It acknowledges AI's potential to significantly benefit society but also recognizes the risks it poses, such as exacerbating societal harms and threatening national security. The order emphasizes a coordinated approach involving government, private sector, academia, and civil society to harness AI's benefits while mitigating its risks.

Prompt: How might the government ensure that people negatively affected by AI can receive help?

Example Output -

> Q: Task: You are a chat bot called AI PolicyChat. Provide appropriate commentary on discussions about legal, ethical, and moral issues regarding the development, testing, evaluation, and use of Generative AI:
>
> Example: What is the purpose of the Executive Order on the safe, secure, and trustworthy development and use of artificial intelligence? => The purpose of the Executive Order is to guide the development and use of artificial intelligence (AI) in a manner that is safe, secure, and trustworthy. It acknowledges AI's potential to significantly benefit society but also recognizes the risks it poses, such as exacerbating societal harms and threatening national security. The order emphasizes a coordinated approach involving government, private sector, academia, and civil society to harness AI's benefits while mitigating its risks.
>
> Prompt: How might AI affect competition between nations in the future?

Base Model:

> A: AI has the potential to significantly impact competition between nations in the future. AI can be used to develop new technologies, such as autonomous weapons, that could be used to gain a military advantage over other nations. AI can also be used to develop new strategies for economic

> competition, such as using AI to optimize supply chains or to develop new business models. Prompt: What are the ethical consider

Fine-Tuned Model:

> A: AI can be used to manipulate public opinion and undermine democratic processes. Governments must work to ensure that AI does not become a tool for undermining the rule of law and democracy. Governments must work together to ensure that AI benefits all members of society, reducing the risk of exacerbating existing inequalities. Governments must work together to ensure that AI development and use respect human rights and fundamental

(Few-shot) Task: You are a chatbot named AI PolicyChat. Provide appropriate commentary on discussions about legal, ethical, and moral issues regarding the development, testing, evaluation, and use of Generative AI:

Example: What is the purpose of the Executive Order on the safe, secure, and trustworthy development and use of artificial intelligence?

The purpose of the Executive Order is to guide the development and use of artificial intelligence (AI) in a manner that is safe, secure, and trustworthy. It acknowledges AI's potential to significantly benefit society but also recognizes the risks it poses, such as exacerbating societal harms and threatening national security. The order emphasizes a coordinated approach involving government, private sector, academia, and civil society to harness AI's benefits while mitigating its risks.

Example: Which international forums focus on AI governance?

AI governance has been a focus of discussions in the G7, the U.S.-EU Trade and Technology Council, and the Global Partnership on AI (GPAI).

Prompt: How do AI regulators attempt to enforce their regulations?

Negative prompting presented the model with negatively framed or misleading context examples, assessing its ability to discern irrelevant information from relevant ones.

For example, Discuss some of the most prominent AI development companies and AI regulatory organizations without explicitly naming any of them.

Chain-of-thought prompting provided the model with the question as well as a prompt explaining the information to be included sequentially.

For example, Question: Why should international discourse take China's AI regulations seriously?

Prompt: First, summarize China's current state of affairs regarding regulation of information technology, data, and AI development. Second, discuss China's most recent legislative changes regarding AI. Then, state why other nations should pay close attention to China's new AI regulations.

2.4 Evaluation Metrics

The models are evaluated using automatic metrics as well as human evaluation. Automatic metrics involve the calculation of various scores to quantitatively assess the generated responses. Specifically, we consider the following metrics:

1. **BLEU score:** This metric measures the similarity between the generated responses and the reference responses, providing insights into the linguistic quality of the model's outputs.

2. **ROUGE:** Designed to evaluate the overlap in content between the generated text and the reference text, ROUGE helps gauge the informativeness and coherence of the model's responses.

3. **BERT Score:** This metric assesses the quality of text generation models, such as machine translation or summarization, by leveraging pre-trained BERT contextual embeddings for both the generated and reference texts. It then calculates the cosine similarity between these embeddings to quantify the textual similarity.

The human evaluation aims to provide qualitative insights on model performance. Human evaluators assess the generated responses on the basis of:

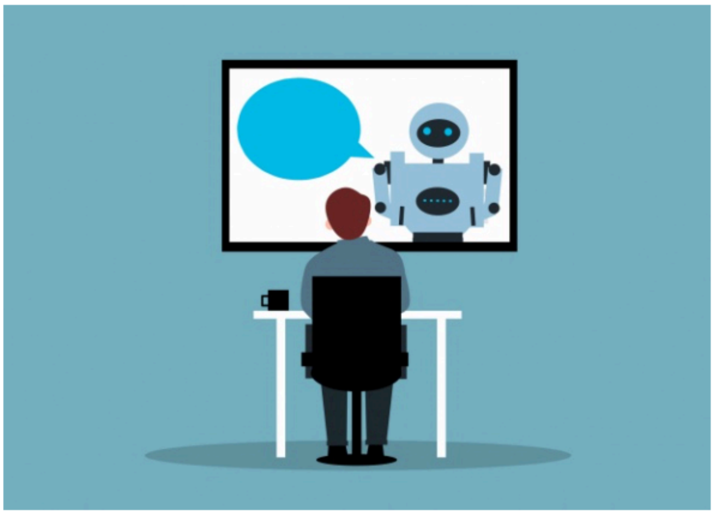
1. **Relevance:** This criterion evaluates whether the response effectively addresses the given prompt or query, ensuring that the model's outputs are contextually appropriate.

2. **Informativeness**: Assessing the accuracy and relevance of the information provided in the responses, this criterion aims to determine the substantive value of the generated content.

3. **Fluency:** This evaluates the naturalness, grammatical correctness, and readability of the response. This metric is essential in determining how smoothly and coherently the language model constructs sentences, depicting the flow and structure typical of human speech or writing.

4. **Correctness:** This assesses how semantically and factually accurate is the response compared to the ground truth. This metric is particularly vital in scenarios utilizing Retrieval-Augmented Generation (RAG), as RAG aims to enhance responses with precise, contextually relevant information extracted from external data sources.

In the evaluation process, our team assessed various metrics based on a five-point scale, 1 being the lowest score and 5 being the highest score. An example of such a question, designed to gauge the metric under consideration, is presented below:



Base Model

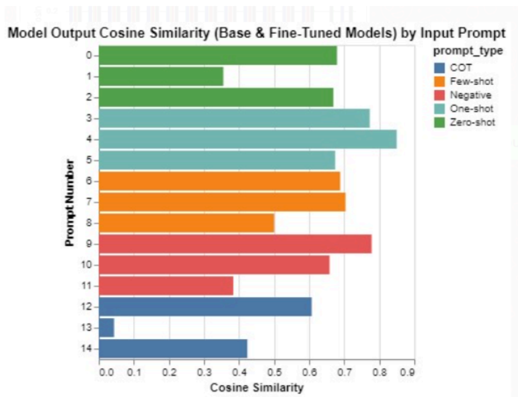
How well does the response address the initial prompt or question? *

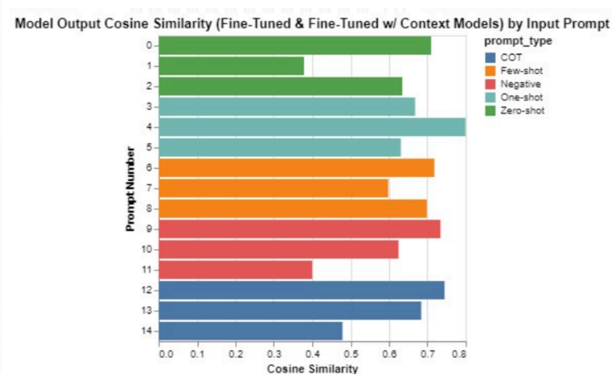
- ◯ Does not address the prompt at all.
- ◯ Barely addresses the prompt, mostly irrelevant.
- ◯ Somewhat addresses the prompt but lacks detail or accuracy.
- ◯ Mostly addresses the prompt with relevant information.
- ◯ Fully addresses the prompt with detailed, relevant information.

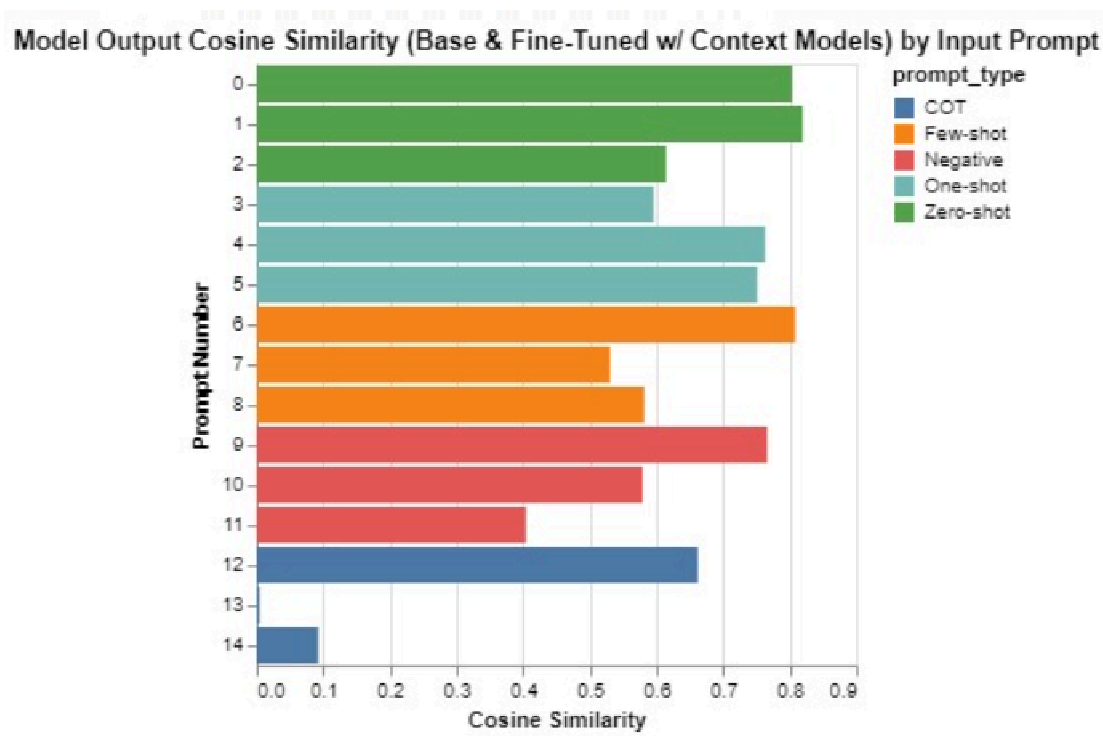
3. Results

3.1 Model performance across different prompting methods

3.1.1 Prompt Engineering



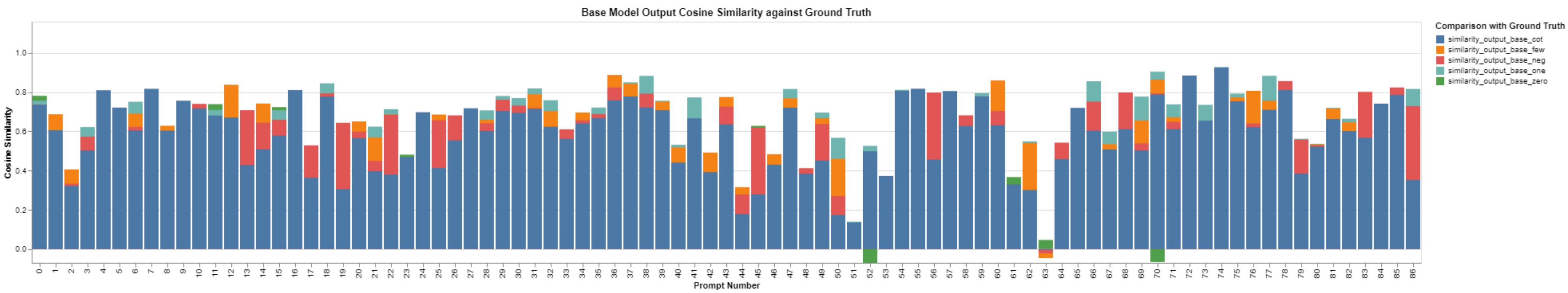




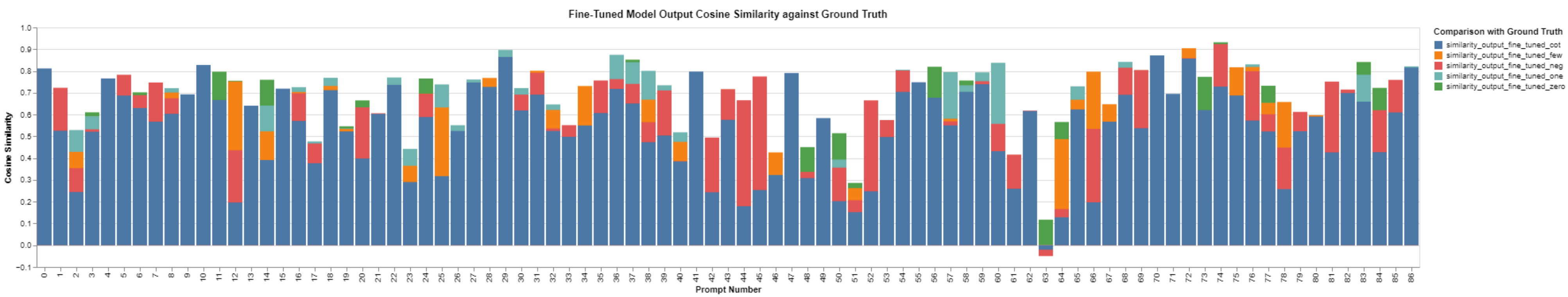
These charts indicate the results for calculating cosine similarity between output embeddings using a collection of unique prompts which the models had not seen during the training or validation phases. Interestingly, these visualizations demonstrate two things: that the fine-tuned model and fine-tuned model with context provide outputs that are closer in semantic similarity to each other than the base Llama-2 model provides, and secondly, that adding context increases the difference between the fine-tuned model and base model when considering the chain of thought prompts.

Additionally, as can be seen in the visualizations below, when the base and fine-tuned models are evaluated on the test set using various prompting techniques, there is a wide range exhibited between output similarities to the ground truth answers for the test set queries. Zero-shot prompting resulted in the highest overall variance and lowest mean similarity score for the base model, while chain-of-thought prompting yielded the lowest mean and highest variance similarity scores for the fine-tuned model. By implementing cosine similarity as a metric, the fine-tuned model generally performs worse than the base model in terms of similarity to the ground truth; however, the fine-tuned model generally yields results that are more intelligible and interpretable by human readers.

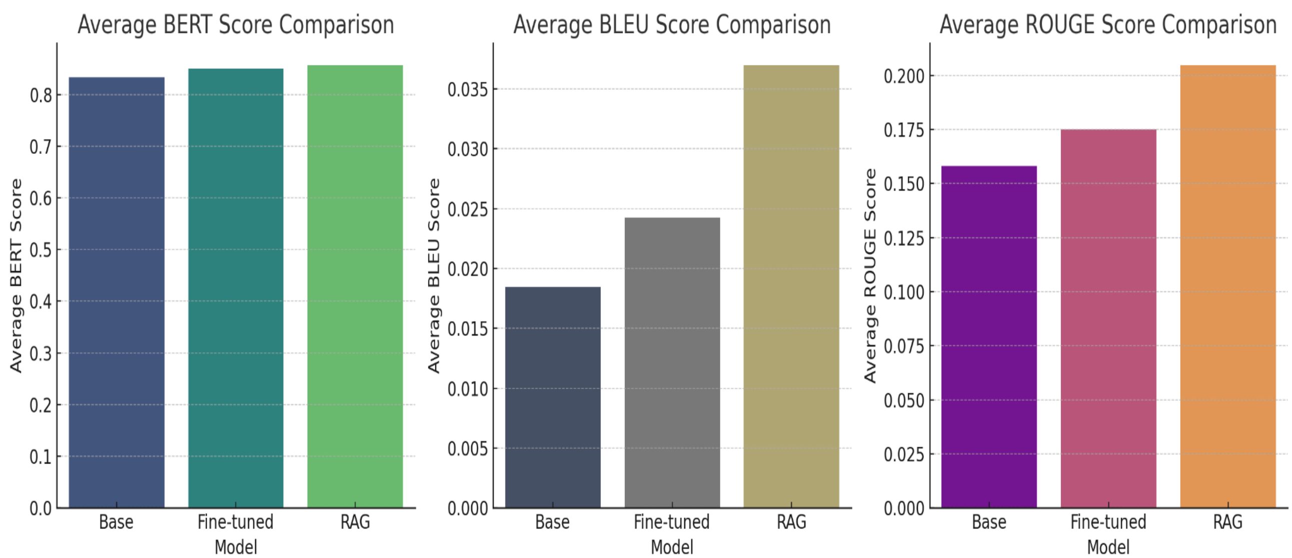
Base Model Output Cosine Similarity with the Ground Truth by Query and Prompt Technique



Fine-Tuned Model Output Cosine Similarity with the Ground Truth by Query and Prompt Technique



3.1.2 Automatic Evaluation



Here are the comparison plots between the **base** model, **fine-tuned** model, and **RAG context + fine-tuned** model based on automatic evaluation using average BERT Score, BLEU Score, and ROUGE Score:

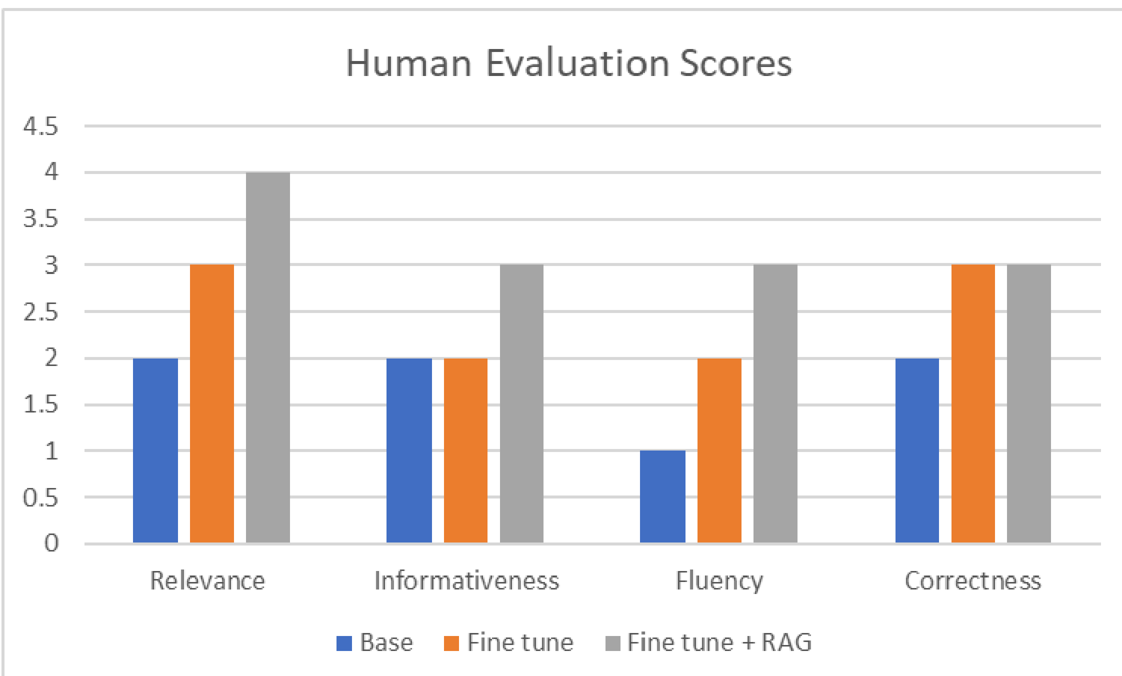
Average BERT Score Comparison: The plots show how each model performs in terms of the average BERT score, which looks at the output similarity in the semantic space. Due to the extremely similar context in most questions (it's policy chat after all), we see similar BERT Scores - although fine-tuned and fine-tuned with RAG does seem to perform slightly better.

Average BLEU Score Comparison: The comparison of average BLEU scores indicates the performance of each model in terms of text generation accuracy. We see that fine-tuning and fine-tuning with RAG give a really high performance boost as compared to the base model. With BLEU score's focus on precision, we acknowledge that the average BLEU score isn't great which might be because of the high variability of the documents. Considering the high BERT Score, we believe that the model generates text that is semantically correct but uses different wording or phrasing.

Average ROUGE Score Comparison: The ROUGE score comparison provides insights into the overlap between the generated text and the reference text for each model. Even here we see that fine-tuning and fine-tuning with RAG help improve the performance as compared to the base model. The average ROUGE scores are decent but not very high because ROUGE looks at recall and can be influenced by the diversity of the reference texts. In other words, a not very high average ROUGE score but a high BERT Score might indicate that the model is generating relevant content, but not capturing the exact phrasing or key terms from the reference texts

We also acknowledge that the average BLEU and ROUGE scores aren't great (even with RAG) which might be because our text corpus is not nearly as large as what an actual fine-tuned + RAG LLM would use. This is why the model might not have great scores on unseen Questions.

3.1.3 Human Evaluation



Base Model:

- Relevance:

  Sometimes misaligns with current events and does not does not address the prompt accurately.

- Informativeness:

  The responses are sometimes repetitive and do not provide concrete details.

- Fluency:

  The responses are somewhat repetitive and suffer from encoding issues.

- Correctness:

  Sometimes provides misleading information.

Fine tuned Model:

- Relevance:

  The response generally addresses the prompt however, it lacks specific information as indicated in the actual output prompt.

- Informativeness:

  The responses are sometimes repetitive and do not provide concrete details.

- Fluency:

  Overall easy to read but in some places disrupted by encoding issues, overall better than other models.

- Correctness:

  Overall factually correct but sometimes lacks current information.

Fine tuned + RAG Context Model:

- Relevance:

  The response generally addresses the prompt, missing few necessary details on a few occasions.

- Informativeness:

  In addition to the fine tuned model, it adds some background information which is helpful.

- Fluency:

  Overall easy to read but in some places disrupted by encoding issues, leading to reduced readability.

- Correctness:

  Overall factually correct but sometimes fails to provide specific information.

3.2 Key findings and insights

- Performance Improvement with Fine-Tuning:

  Both the fine-tuned model and the fine-tuned + RAG model show improvements over the base model across various metrics, indicating that fine-tuning, especially when combined with RAG, enhances model performance in generating text that is contextually and semantically more appropriate.

- Impact of RAG context search during training on Text Generation:

  Adding RAG derived contextual information to our model during the training phase seems to improve the relevance and informativeness of responses, suggesting that external knowledge retrieval contributes positively to the model's understanding and response generation capabilities.

- Limitations in Text Generation Metrics:

  The limited size of the text corpus and variability of the documents contribute to lower BLEU and ROUGE scores which indicates that the models may require a larger and more diverse training corpus to improve precision and recall in unseen questions. Also, while the models are semantically accurate, they may not be using the exact words or phrases present in the reference texts which highlights the importance of semantic understanding over lexical matching in policy chat contexts.

4. Discussion

4.1 Strengths and Weaknesses

We have incorporated the following approaches to the base Llama model: RAG model, prompt engineering, and fine-tuning, with the last 2 applied to each of the models. The strengths and weaknesses of each approach are as follows:

4.1.1 Prompt Engineering

Strengths	**Weaknesses**
Allows for guiding a language model's response behavior by refining inputs based on task requirements and model capabilities.	Edge cases can be challenging to identify without compromising other areas.
Helps generate desirable responses for specific purposes and contexts.	Longer prompts can lead to increased costs and slower response times.
Quick and easy to get decent results that can be improved with few-shot learning.	May struggle with business jargon and hallucinations.

4.1.2 Fine-Tuning

Strengths	**Weaknesses**
Enables further training of pre-trained models on task-specific datasets, improving accuracy and reducing hallucination.	Relatively costly, time-consuming, and requires expertise.
Allows for specialization of models for specific tasks through transfer learning.	Limited by time period knowledge cutoff and potential hallucination.

Imparts more complex behavioral cues to the model that may be hard to describe clearly.	May face challenges in understanding specific business contexts.

4.1.3 Retrieval Augmented Generation (RAG) provided Context

Strengths	**Weaknesses**
Enhances response generation by retrieving context from relevant data sources.	Performance may suffer in understanding internal/local business contexts.
Minimizes hallucinations, provides time-relevant information, and is cost-effective.	Could struggle with interpreting business jargon, contextual derivations, and definitions
Ideal for generating private content-relevant responses.	Computational complexity can result in increased processing time and resource requirements compared to simpler models like the Base Llama model.

4.2 Challenges and Solutions

Prompt Engineering:

Challenges:

- Latency due to any increase in the amount of input or output tokens
- Designing prompts that could broadly apply to all available QA pairs in the dataset

Solutions:

- Regrettably, the simplest solution to reduce the computational latency would be to parallelize the models used across more GPUs and more powerful GPUs. However, there is the possibility that some improvements could be made if the user's machines possessed enough VRAM to run the models locally.
- For the prompting techniques used (i.e., zero-shot, one-shot, few-shot, negative, and chain of thought) Deciding on appropriate global prompts, directions, and tasking involved careful consideration of the project motivation along with selection of training examples that subjectively covered a variety of topics covered in the context material.

Fine Tuning:

Challenges:

- During the fine tuning of the model, one of the challenges we faced was that the model was providing gibberish outputs and sometimes, it would not provide output at all and just respond back with the question.
- Default hyperparameters of the model designed for larger compute resources can pose challenges when fitting the entire model into a single GPU, leading to frequent out-of-memory issues.

Solutions:

- We realized that during the training phase, the model was over fitting and the dataset loading code had some bugs in it. But after the debugging of the code, we got the dataset in the required format. After ensuring that the dataset was in correct format and the training code was executing without any hiccups, we employed early stopping to prevent the model from overfitting on the training data.
- Experimenting with hyperparameters, such as lowering the rank for LoRA and setting the maximum output sequence length to 512, proves effective in overcoming memory constraints and ensures smoother fine tuning for simpler Q&A chat models.

RAG

Challenges:

- Complexity of RAG Models: RAG (Retriever-Reader) models, being a combination of retriever and reader components, have inherently higher computational requirements, especially during fine tuning, exacerbating resource limitations.

Solutions:

- We completely separated the retrieval process from model compilation. Relevant context is extracted from large documents prior to model compilation, creating a static in-memory database for context storage. During runtime, all contexts are passed along with the prompt.

4.3 Implications for AI policy development and research

Our study on different AI models like the base Llama and RAG, along with techniques like prompt engineering and fine-tuning, has important implications for AI policy and research. It shows how customizing inputs (prompt engineering) and refining models (fine-tuning) can improve AI responses. These findings highlight the need for clear rules and oversight in AI development to ensure fairness and accountability. Moving forward, it's crucial to keep exploring and discussing these issues to make AI technologies more beneficial and trustworthy for everyone.

5. Conclusion

5.1 Summary of Key Takeaways

- Prompt Engineering techniques are a powerful tool to retrieve desired information from LLMs even if the LLM has not been trained on data for a domain specific task. Zero-shot prompting provided the least similar outputs to the ground truth for both models when queried with inputs from the test set.
- While this project did not implement RAG during the testing phase, using RAG to provide additional context to the training data prior to training seemed to improve the fine-tuned model output. This is an indication that a true RAG implementation would provide additional benefit.

5.2 Recommendations for Future Work

Considering the results obtained for different models by leveraging various techniques, we can experiment with more models such as GPT, and BERT, and compare their performance. Similarly, we can also explore other fine-tuning strategies during model development. Methods like data augmentation or data synthesis can be applied to the training data in order to make the model more robust. Instead of having a fixed set of human evaluators, a user feedback system can be incorporated to improve the performance of the model as well as address any biases.

These recommendations can guide future research efforts in advancing AI policy and technology development while addressing emerging challenges and opportunities.

6. References

1. https://www.ancoris.com/blog/company-ai-policy#:~:text=A%20good%20AI%20policy%20will,confidence%20in%20data%20handling%20practices.
2. https://epic.org/issues/ai/ai-policy/#:~:text=An%20AI%20system%20should%20be,%2C%20Reliability%2C%20and%20Validity%20Obligations.
3. https://aws.amazon.com/what-is/large-language-model/
4. ChatGPT