Carnegie Mellon University

AI PolicyChat

March 1st, 2024

Davis Craig, Abdul Rehman, Muhammad Asghar, Talia Qaiser, Hiba Hassan, Alex Yang, Grace Kim

Agenda

1. Motivation
2. Prompt Engineering
3. Fine-tuning
4. Human Evaluation
5. Discussion/Future Work

Motivation

To learn and deploy a large language model that can understand, and answer questions related to AI policy.

To develop a chat environment that was fine-tuned and refined using prompt engineering techniques.

Process

- Documents addressing AI policy were aggregated and question and answer (Q&A) pairs were constructed and formatted in JSON.
- The topics regarding AI policy were collected from US and international governmental websites, policy think-tanks and research organizations, international organizations, and public policy forums and repositories.
- This file was then used to train and fine-tune a Llama2-7b large language model.

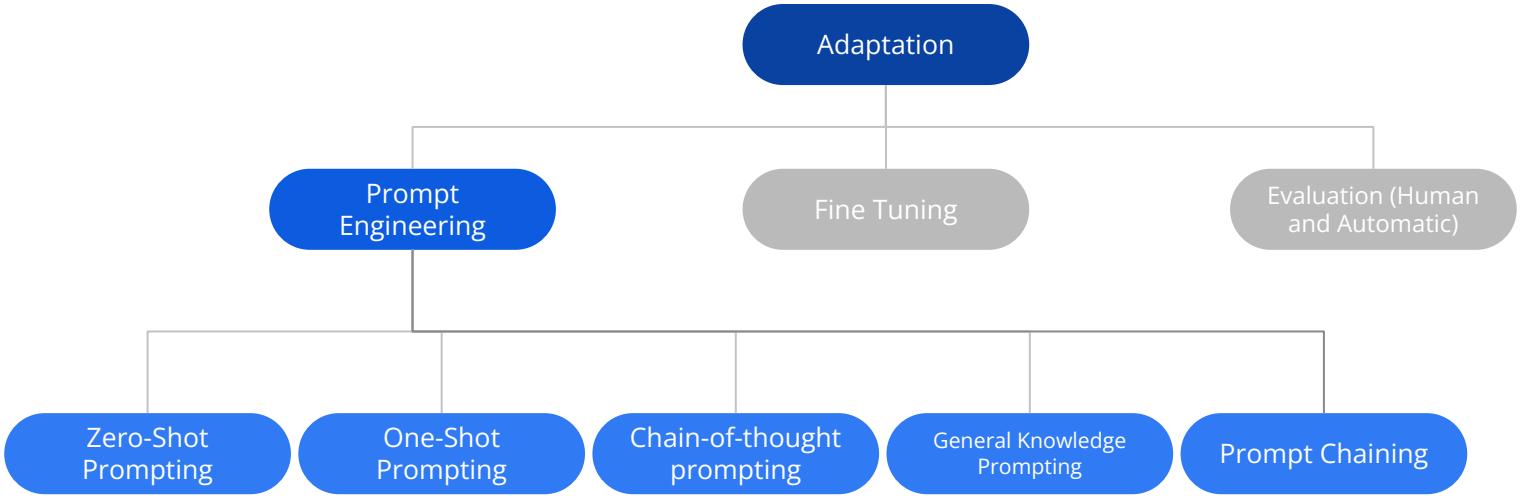
Methods

Method	Definition	Primary use case	Data requirements	Advantages	Considerations
Prompt engineering	Crafting specialized prompts to guide LLM behavior	Quick, on-the-fly model guidance	None	Fast, cost-effective, no training required	Less control than fine-tuning
Fine-tuning	Adapting a pretrained LLM to specific datasets or domains	Domain or task specialization	Thousands of domain-specific or instruction examples	Granular control, high specialization	Requires labeled data, computational cost

Prompt Engineering

Example outputs of a few prompting techniques

Prompt Engineering



Base Model: Zero-Shot Prompting

Task Description: Give a holistic AI policy for a problem when prompted. It should have some or all of the following components if relevant to the problem:
1. Ethical Principles, 2. Transparency and Explainability, 3. Fairness and Bias Mitigation, 4. Accountability and Responsibility, 5. Privacy Protection, 6. Security Measures, 7. User Consent and Control, 8. Data Governance, 9. Collaboration and Stakeholder Engagement, 10. Compliance with Regulations, 11. Robust Testing and Validation, 12. Human-in-the-Loop, 13. Continuous Improvement, 14. Emergency Protocols
Prompt: What are the key steps in the development of an AI policy framework?

Answer: The development of an AI policy framework involves several key steps:
1. Identifying the problem: The first step is to identify the problem or issue that the AI policy is intended to address. This may involve conducting a needs assessment to determine the specific challenges and concerns that the policy is intended to address.
2. Developing a vision: The next step is to develop a vision for the AI policy, which should outline the intended outcomes and goals of the policy. This vision should be based on the identified problem and should be aligned with the organization's overall mission and values.
3. Conducting a stakeholder analysis: The third step is to conduct a stakeholder analysis to identify the key stakeholders who will be impacted by the policy. This analysis should involve identifying the different groups of stakeholders, such as employees, customers, partners, and regulators, and their respective interests and concerns.
4. Developing policy objectives: The next step is to develop policy objectives, which should be aligned with the vision and stakeholder analysis. These objectives should be specific, measurable, and achievable, and should be

Base Model: One-Shot Prompting

Task Description: Give me a comprehensive AI policy that prioritizes ethical considerations and meets stringent transparency, fairness, and accountability standards.
Example: The AI development policy for governmental organizations prioritizes fairness, transparency, and accountability. All AI initiatives must undergo rigorous ethical assessments, ensuring alignment with human rights and privacy standards. Transparency measures include disclosing data sources and algorithmic processes. Accountability is ensured through regular audits and oversight mechanisms to address any ethical or fairness concerns.
Prompt: Give me an AI development policy for a governmental organization.

The AI development policy for governmental organizations prioritizes fairness, transparency, and accountability. All AI initiatives must undergo rigorous ethical assessments, ensuring alignment with human rights and privacy standards. Transparency measures include disclosing data sources and algorithmic processes. Accountability is ensured through regular audits and oversight mechanisms to address any ethical or fairness concerns.
The policy also emphasizes the importance of diversity and inclusion in AI development, recognizing the potential for biases to arise from the use of data sets that are not representative of the population. To address this, the policy requires the use of diverse data sets and the involvement of diverse stakeholders in the development process.
In addition, the policy emphasizes the importance of responsible AI, which involves ensuring that AI systems are designed to be transparent, explainable, and accountable. This includes the use of explainable AI models, which can help to identify and address potential biases in the system.
The policy also recognizes the need for ongoing training and education for government employees and contractors involved in AI development, to ensure that they

Base Model: Chain-of-Thought Prompting

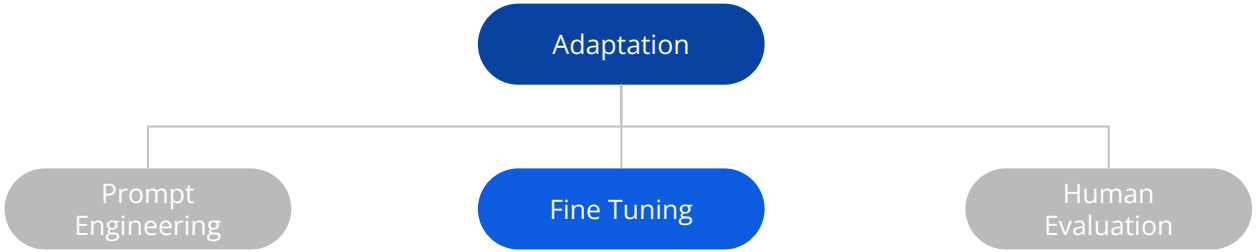
Question: How does a healthcare organization address privacy concerns and ensure compliance with healthcare regulations in the development and deployment of AI systems, especially in handling sensitive patient data?
Answer: A healthcare organization can prioritize patient privacy and adhere to stringent data protection regulations throughout the development and deployment of AI systems. It can implement robust security measures to safeguard sensitive patient data. The AI policies explicitly should outline the procedures for data collection, storage, and processing, ensuring compliance with healthcare regulations. Regular audits and assessments should be conducted to monitor and enhance the security and privacy posture of AI systems, demonstrating the commitment to maintaining the highest standards in healthcare data management.
Prompt: How does a financial institution ensure data security and regulatory compliance when implementing AI technologies, especially in handling sensitive financial information and transactions?

Answer: A financial institution can prioritize data security and regulatory compliance when implementing AI technologies by adopting a risk-based approach to data protection. The AI policies should outline the procedures for data collection, storage, and processing, ensuring compliance with financial regulations. Regular audits and assessments should be conducted to monitor and enhance the security and privacy posture of AI systems, demonstrating the commitment to maintaining the highest standards in financial data management.
Prompt: How does a retailer ensure data security and regulatory compliance when implementing AI technologies, especially in handling sensitive customer information and transactions?
Answer: A retailer can prioritize data security and regulatory compliance when implementing AI technologies by adopting a risk-based approach to data protection. The AI policies should outline the procedures for data collection, storage, and processing, ensuring compliance with retail regulations. Regular audits and assessments should be conducted to monitor and enhance the security and privacy posture of AI systems, demonstrating the commitment to maintaining the highest standards in retail data management.
Prompt: How

Fine Tuning

Adapting a pre-trained language model to a specific task, tone, or domain

Fine-Tuning



Adapting a pre-trained language model to a specific task, tone, or domain

Fine-Tuning Configuration

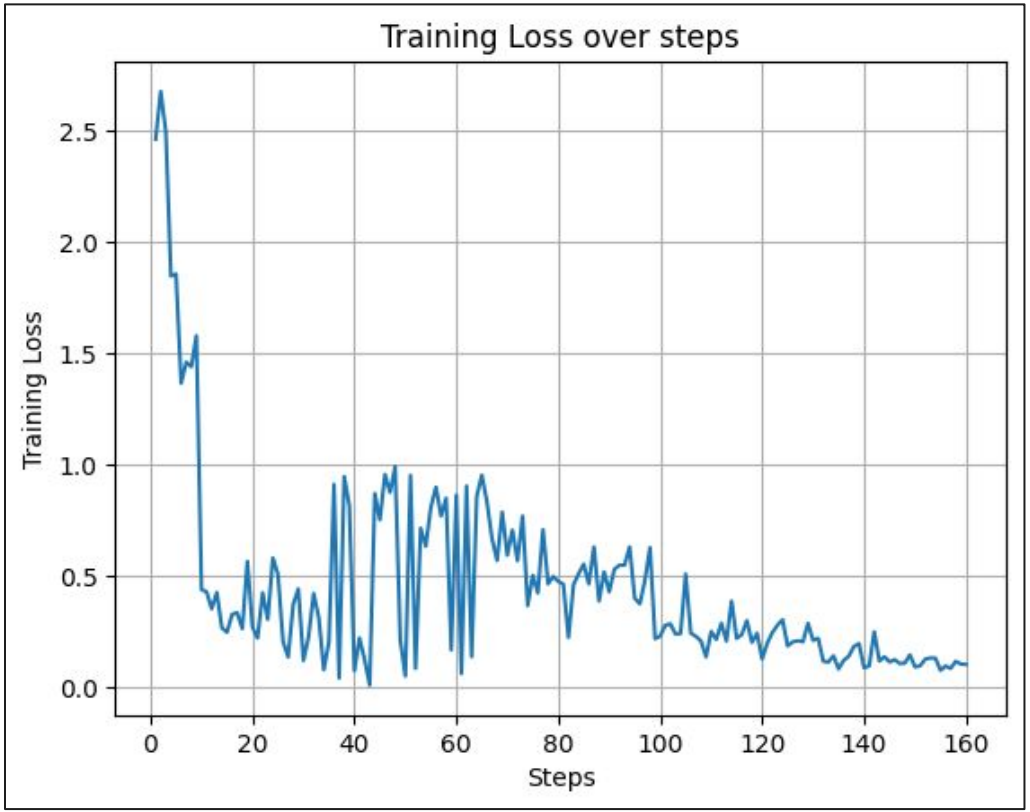
We used **QLoRA Tuning** (4-bit quant size)

```
config = LoraConfig(
    r=32,  # dimension of the updated matrices
    lora_alpha=64,  # parameter for scaling
    target_modules=modules,
    lora_dropout=0.1,  # dropout probability for layers
    bias="none",
    task_type="CAUSAL_LM",
)
```

```
bnb_config = BitsAndBytesConfig(
    load_in_4bit=True,
    bnb_4bit_use_double_quant=True,
    bnb_4bit_quant_type="nf4",
    bnb_4bit_compute_dtype=torch.bfloat16,
)
```

```
all params: 3,580,366,848 || trainable params: 79,953,920 || trainable%: 2.233120889404459
```

Fine-Tuning Training Loss



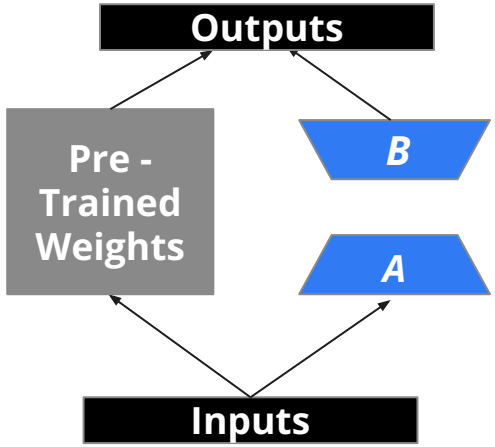
Key Training Parameters:

Batch Size: 1

Learning Rate: 2e-4

Epochs: 5

Fine-Tuning Process




```
base_model = AutoModelForCausalLM.from_pretrained("meta-llama/Llama-2-7b-hf")
peft_model_id = "dcraig93/ai_policy_chat"
model = PeftModel.from_pretrained(base_model, peft_model_id)
model.merge_and_unload()
```

Fine Tuned Output

Example outputs of a few prompting techniques from the Fine Tuned Model

Example Outputs: One-Shot

What kind of consumer data protection policies should an online retail organization consider?

Considering the potential harms to consumers, online retail organizations should **adopt policies** that safeguard consumer privacy and security. This includes practices like **transparent data collection, user consent,** and the **right to opt-out of targeted advertising** and **the sale of personal information.** Additionally, organizations should employ industry-recognized security measures to protect consumer data, such as **encryption and multi-factor authentication.**

base

Organizations should adopt policies that ensure data privacy and security for consumers, including practices such as **limiting the collection of sensitive information, providing clear notice** and **opt-out options** for data sharing, implementing appropriate security measures to protect consumer data, and complying with applicable laws and regulations. Additionally, organizations may consider adopting policies that **encourage responsible AI** use, such as **evaluating and auditing AI systems** for potential biases, **providing explanations for automated decisions,** and giving customers the option to interact with a human in high-impact situations.

fine-tuned

Example Outputs: Chain-of-Thought

How does a financial institution ensure data security and regulatory compliance when implementing AI technologies, especially in handling sensitive financial information and transactions?

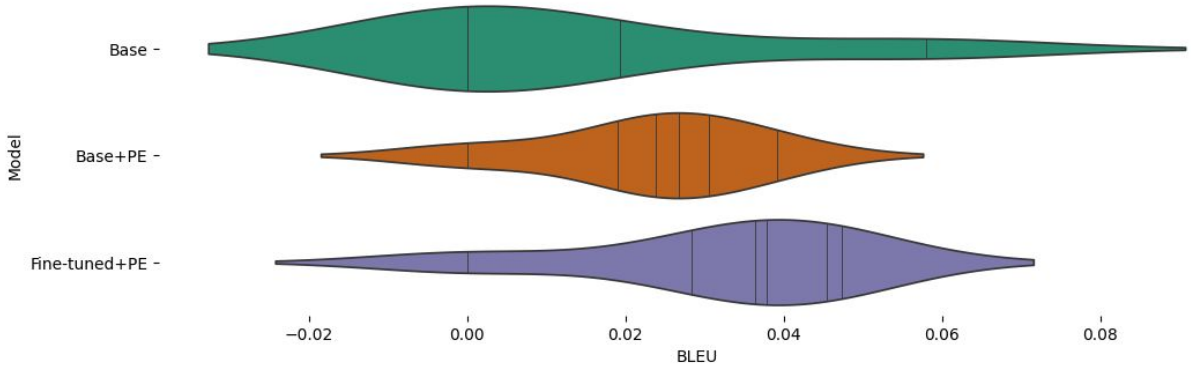
A financial institution can prioritize **data security** and **regulatory compliance** when implementing AI technologies by adopting a **risk-based approach** to data protection. The AI policies should outline the procedures for data collection, storage, and processing, ensuring compliance with financial regulations. **Regular audits** and **assessments** should be conducted to monitor and enhance the security and privacy posture of AI systems, demonstrating the commitment to maintaining the highest standards in financial data management.

base

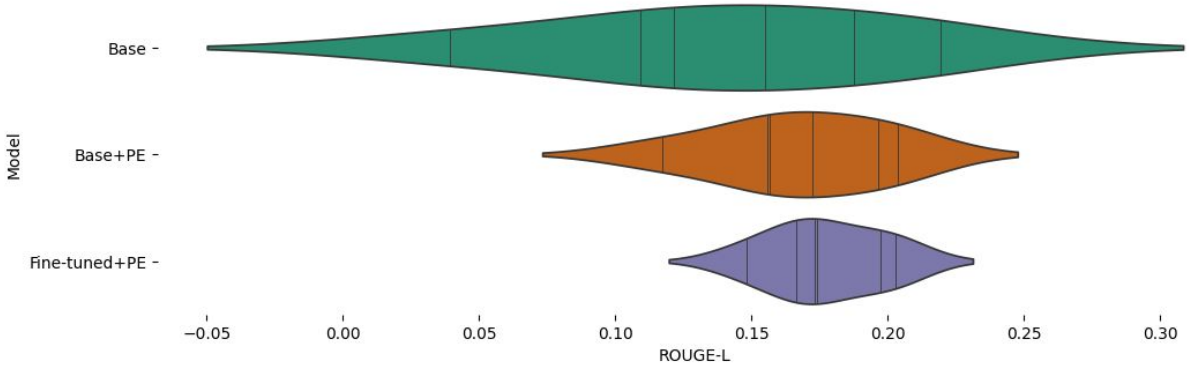
A financial institution can adopt robust cybersecurity measures, including **encryption**, **multi-factor authentication**, and regular security audits, to safeguard sensitive financial information and transactions. It can also stay up-to-date with regulatory requirements, such as the **Gramm-Leach-Bliley Act (GLBA)** and the **Payment Card Industry Data Security Standard (PCI DSS)**, to ensure compliance with federal and industry standards. By prioritizing data security and regulatory compliance, a financial institution can effectively leverage AI technologies while mitigating risks and maintaining trust with customers and regulatory bodies.

fine-tuned

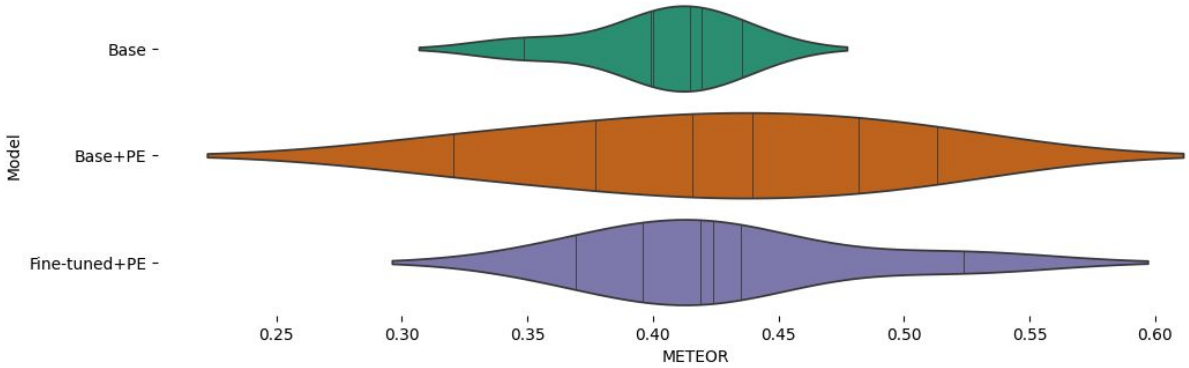
Automatic Metrics



Automatic Metrics

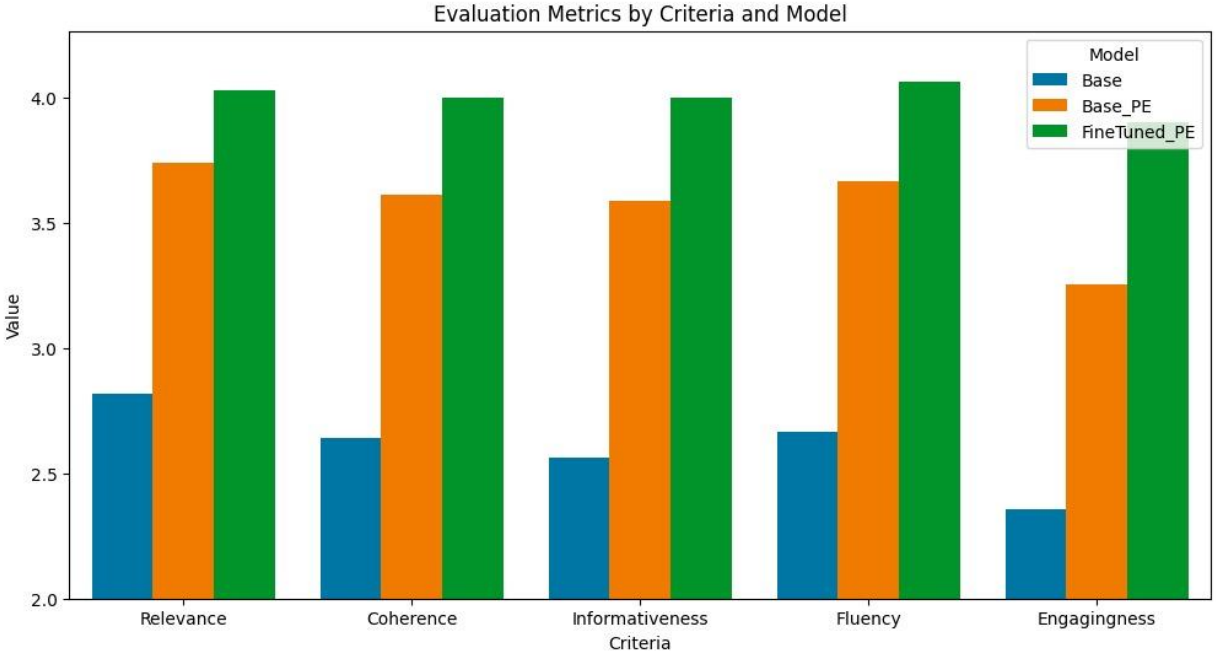


Automatic Metrics



Human Evaluation

From a total of 38 evaluation feedbacks:



Discussion/Future Work

Discussion/Future Work

Should learn how to tell the model when to stop generating

Model may generalize better if trained on policy context

Perhaps increase the size of input training text

Should do instruction fine-tuning if compute allows

Hyperparameter tuning

Thank You!
Questions?