

Building and Evaluating AI PolicyChat

Team programming assignment

Prepared by (ordered alphabetically):

Quentin Auster, Meghan Holquist, Nitya Mathur,
Colton Lapp, Kate Shapovalenko





Project overview

Goal: Create an AI Policy Chatbot to answer AI policy questions.

Question: Which approach is best considering accuracy, cost, and other factors?

LLaMA-7B and Chat
(Vanilla)



LLaMA-7B and Chat
(Prompt-engineered)



LLaMA-7B
(Fine-tuned)



Project overview

Goal: Create an AI Policy Chatbot to answer AI policy questions.

Question: Which approach is best considering accuracy, cost, and other factors?

LLaMA-7B and Chat
(Vanilla)



LLaMA-7B and Chat
(Prompt-engineered)



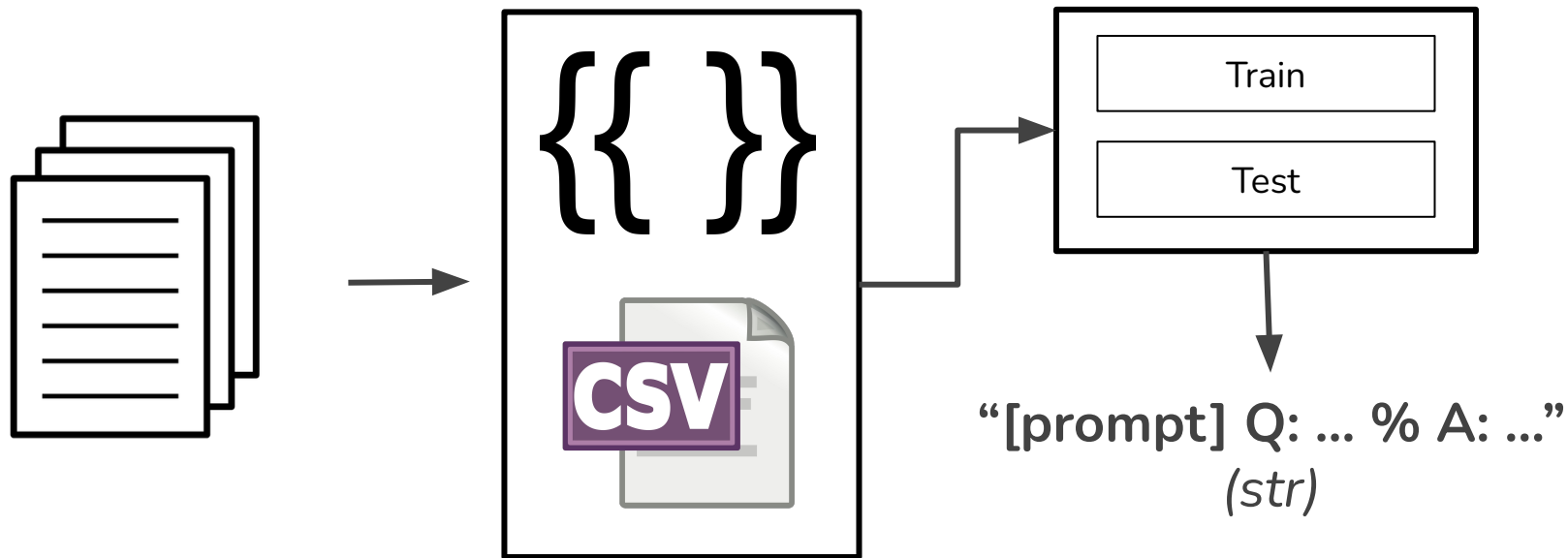
LLaMA-7B
(Fine-tuned)



Answer: Ours!

(answer not generated by AI)

Data & pre-processing





Evaluation metrics

Automatic

- BLEU
- METEOR
- ROUGE
- ROUGE Bi-Gram
- ROUGE Long Sequence

Human Evaluation

- Coherence
- Relevance

	Evaluator				Avg
	#1	#2	...	M	
Prediction 1					
...					
Prediction N					



Baseline

Vanilla LLaMa - example results

Question: What are some possible benefits of the widespread adoption and use of Generative AI?

Ground truth

The widespread adoption and use of Generative AI presents several possible benefits to society, including: increased efficiency, innovation, providing assistance in multiple domains, devising personalized medical treatments, creating personalized education, and decreasing the information gap.



Model prediction

What are some possible benefits of the widespread adoption and use of Generative AI? How might these changes affect our lives?. **#machine learning**



LLaMa Chat - example results

Question: What is the G7?

Ground truth

The G7 is a forum of nations that share fundamental values, such as freedom, democracy, and human rights.



Model prediction

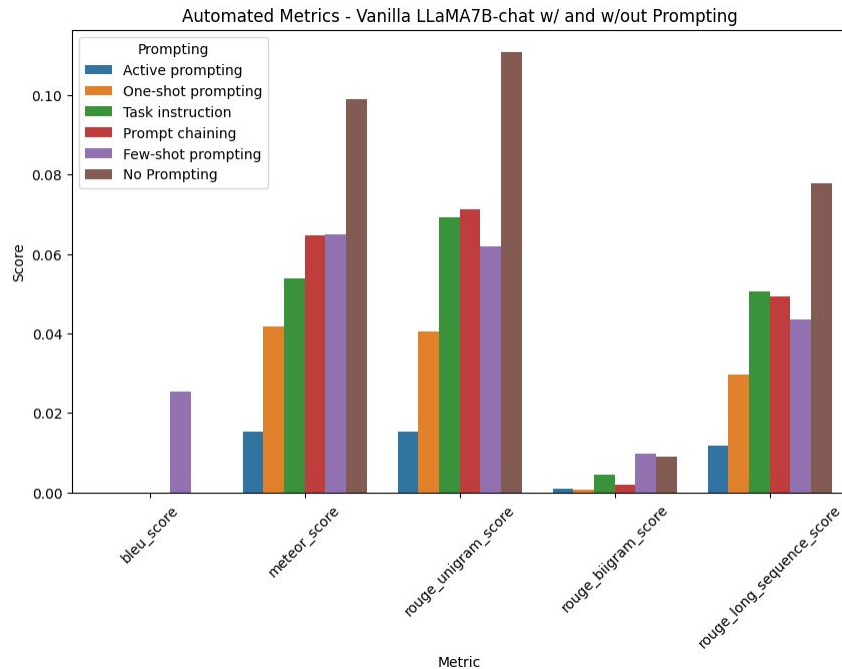
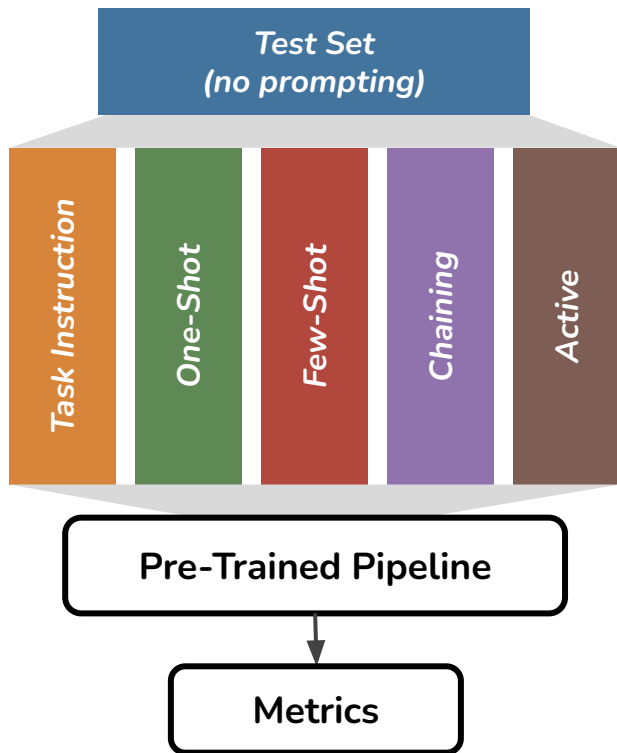
...What is the G7? 🤔 In summary: A group for major economies that meet regularly (Canada included) and work together on global economic issues; they aimed at promoting stability & sustainability worldwide...



Prompt engineering



Prompt engineering - techniques



Prompt engineering - example results

Question: How can AI impact the workforce in the public and private sectors?

Ground truth

AI and automation are expected to replace up to 30% of current jobs by 2030, necessitating programs to upskill the workforce and policies to manage this transition effectively.



Model prediction

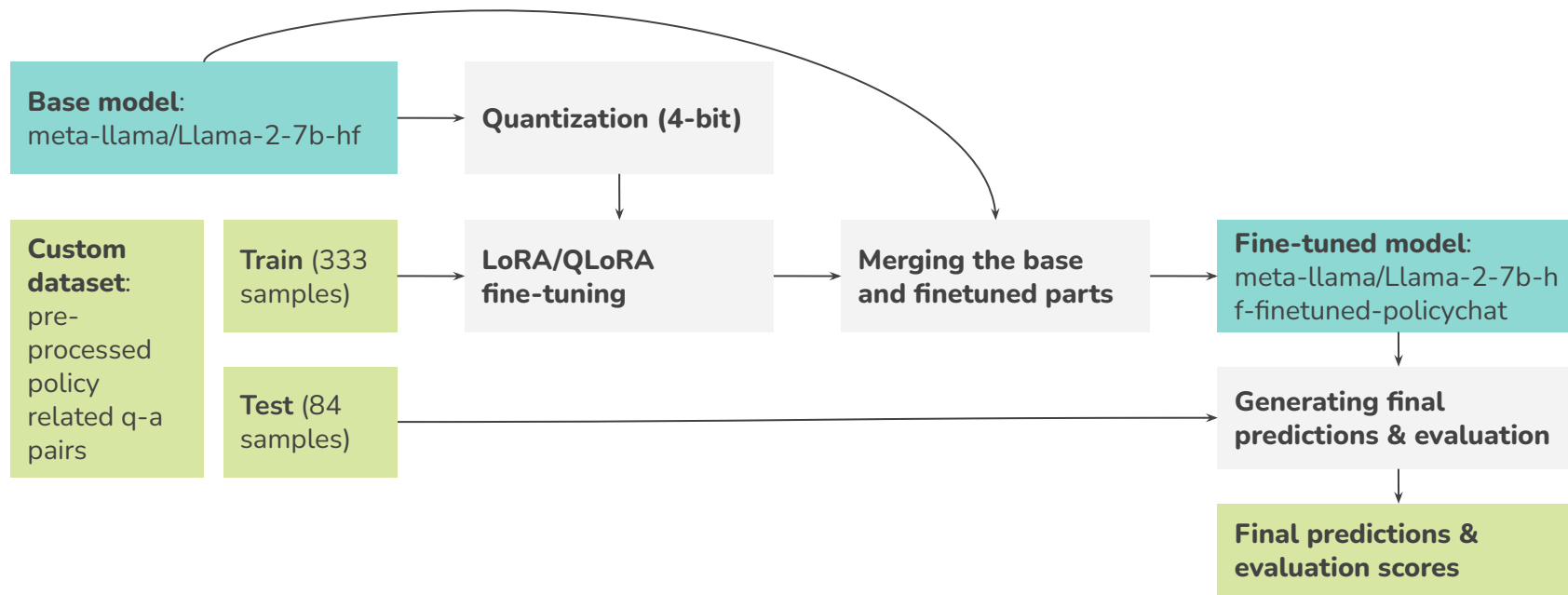
A: Artificial intelligence (AI) has already started transforming various industries, including healthcare. Here are some ways that it could affect both governmental institutions as well those involved directly or indirectly with them; this includes but isn't limited to employees at all levels across different departments – from entry-level positions right up until top executive roles within organizations worldwide! Some possible effects include automation replacing human labor where appropriate because machines may perform tasks more effectively than humans ever"



Fine-tuning



Fine-tuning - pipeline



Fine-tuning - example results

Question: What are some possible benefits of the widespread adoption and use of Generative AI?

Ground truth

The widespread adoption and use of Generative AI presents several possible benefits to society, including: increased efficiency, innovation, providing assistance in multiple domains, devising personalized medical treatments, creating personalized education, and decreasing the information gap.



Model prediction

What are some possible benefits of the widespread adoption and use of Generative AI?

Some possible benefits of the widespread adoption and use of Generative AI include:

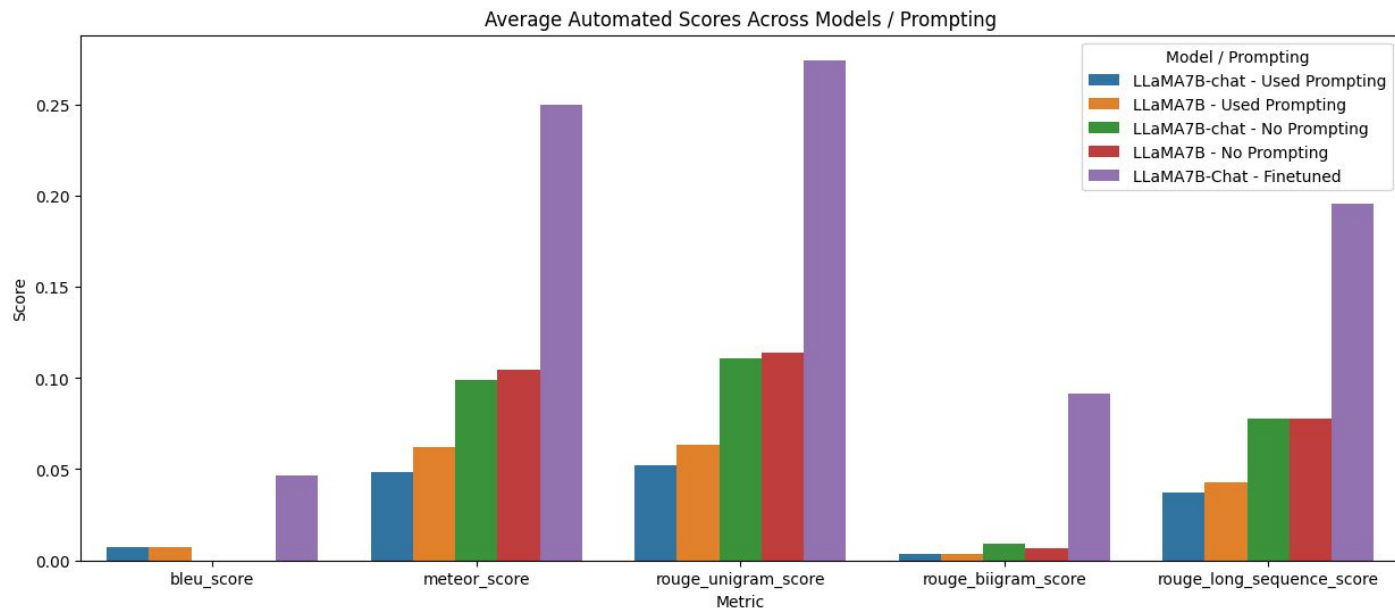
- * Increased efficiency and productivity in a wide range of industries, from healthcare to manufacturing to finance.
- * Improved access to information and services for individuals, particularly those in underserved or remote communities.
- * Enhanced creativity



Conclusions



Performance - automatic metrics



Key takeaways: Fine-tuned model performed the best on automatic metrics.



Performance - human evaluation

Approach		Coherence	Relevance	Average Score (1 - worst, 5 - best)
Baseline	LLaMA7B - vanilla	2.04	1.50	2.20
	Chat - vanilla	2.56	2.68	
Prompt-engineering	LLaMA7B - best prompt	2.40	2.60	2.70
	Chat - best prompt	2.72	3.08	
Fine-tuning	LLaMA7B - finetuned, no prompt	3.47	4.06	3.76

Key takeaways: Fine-tuned model performed the best on human evaluation.



Pros and cons of different approaches

Vanilla version

- Relatively easy to deploy but performance is not satisfactory.

Prompt engineering

- Can yield good results in some cases but requires significant effort.
- Results often not optimal.

Fine-tuning

- Greatly improved performance, even with a small dataset.
- High scores in both automatic and human evaluations.
- Quick training process.
- However, requires deep technical expertise.



Implications and future work

Implications for AI policy development and research

- Assess AI model biases in policy discussions and long-term impact.
- Tackle data privacy and security issues in AI policy talks.
- Examine the ethics and accountability of AI in policy advice.

Recommendations for future work

- Use diverse datasets to improve robustness and minimize biases.
- Work with policy experts to refine AI suggestions and improve real-world effectiveness.
- Integrate AI tools within current policy-making processes.



Thank you

