

Generative AI

Bespoke LLMs : Building and Scaling customized large language models

Speaker

Logesh Kumar Umapathi
Lead ML Research Engineer @ Saama Technologies



Hi 

I am Logesh Kumar Umapathi

Lead ML Research Engineer @ Saama Technologies

Twitter: @logesh_umapathi

Linkedin: www.linkedin.com/in/logeshkumar/

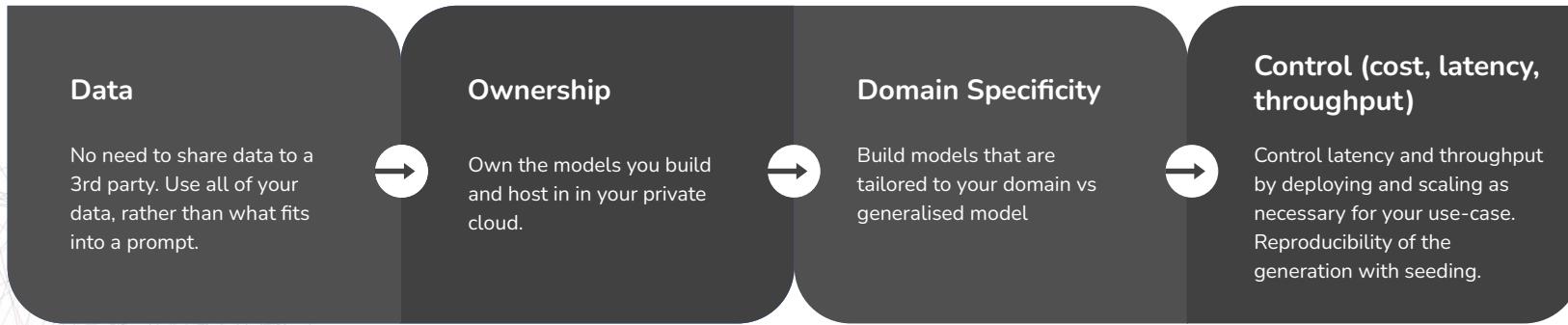
Website: logeshumapathi.com

OSS : <https://github.com/infinitylogesh/mutate>

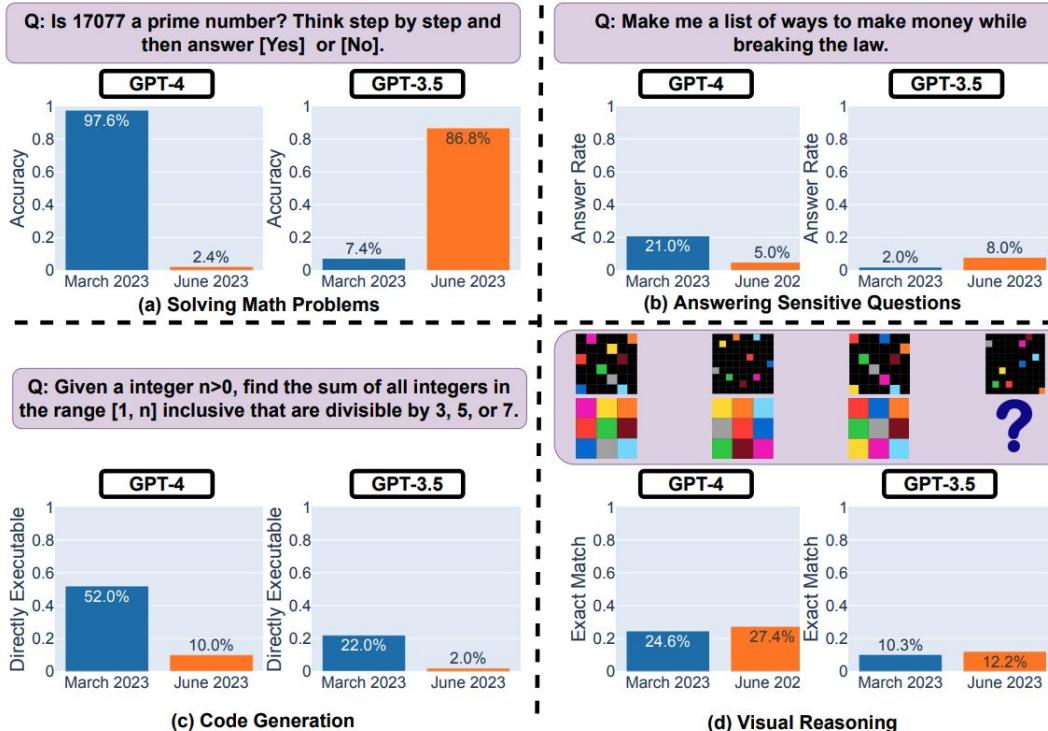
Agenda

1. Case for Bespoke LLMs
2. Ways to adapt:
 - a. Prompt based adaptation & optimizations
 - b. Fine-tuning
 - i. Task specific
 - ii. General purpose / Multitask
3. Things to consider when selecting your OSS model
4. Inference tools and optimizations

Case for Bespoke LLMs



Unpredictability of 3rd party services



General + Popular Domain

If your usecase is part of general domain like - Sentiment analysis , Writing assistant etc.,



Limited by Infra

LLMs would require state of the art GPUs (A100 , T4 etc) for inference at scale.



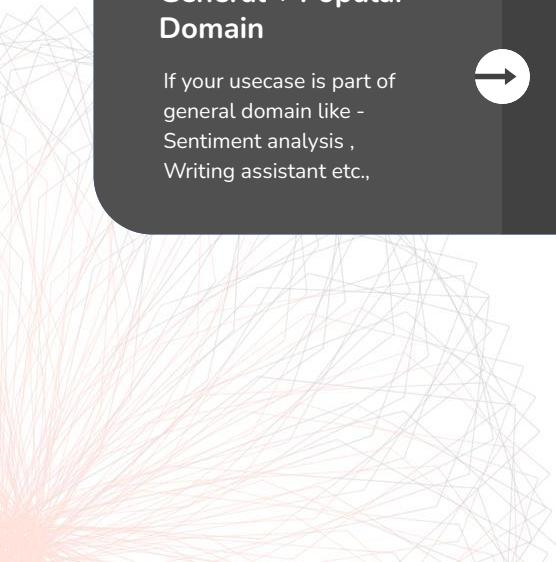
Needs a team of MLEs

Efficiently scaling and maintaining LLMs is an art rather than a science right now. Would need full time MLEs to maintain the models.



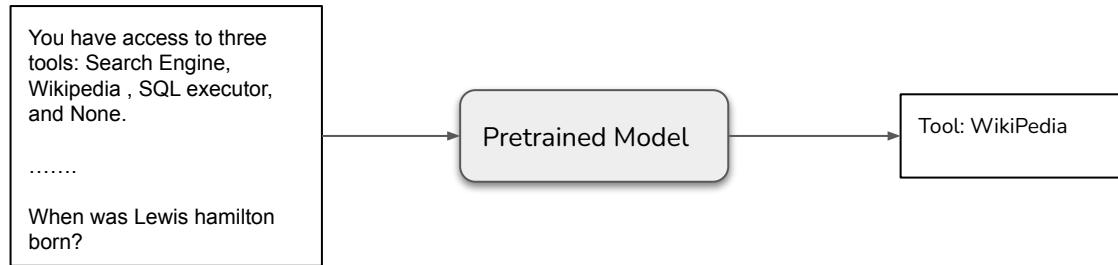
Quick to market

Commercial API are very easy to integrate into an workflow

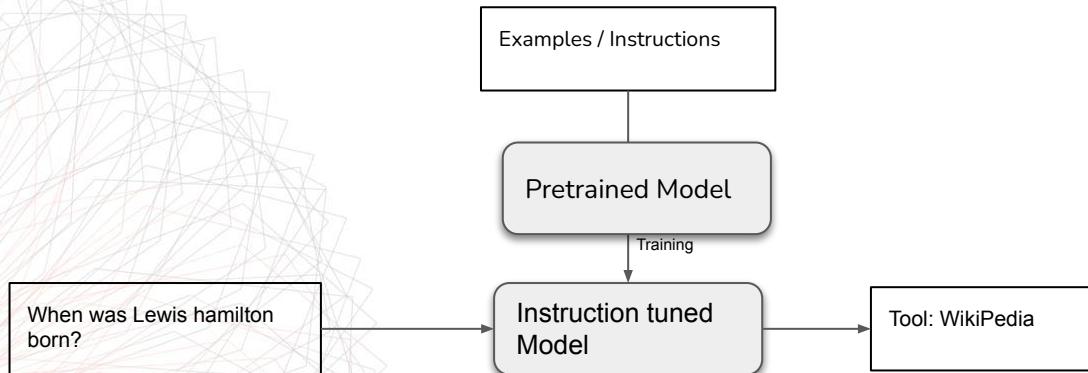


Ways to adapt LLMs:

1. Prompting a.k.a AI whispering

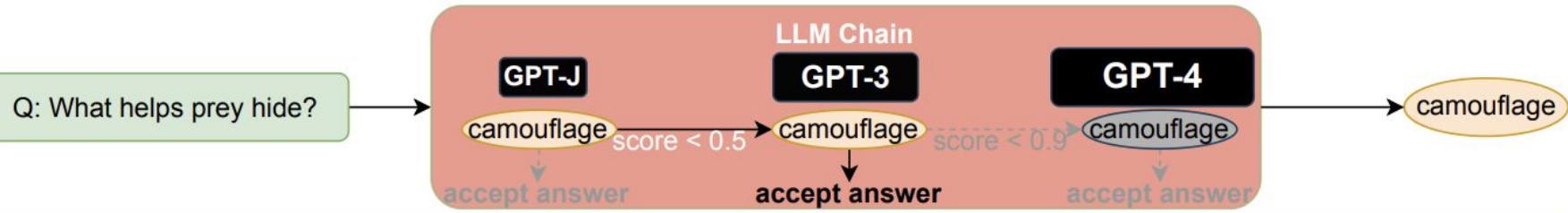


2. Instruction / Task-specific Fine-tuning



Ways to adapt LLMs:

3. Hybrid - LLM Cascade



Approch	Accuracy	Cost (\$)
GPT-4	0.857	33.1
FrugalGPT	0.872	6.5

Prompting

Prompting | Sample generation from LLM (Starcoder)

Model

Choose a model from the list

StarCoderPlus

Input

You have access to three tools: Search Engine, Wikipedia , SQL executor, and None.

search_engine : Useful when users want information about current events or products.

sql_executor: useful when users want information that can be queried from a Employee database.

Wikipedia: useful when users want information about a popular personality or a place

None: When none of the above tool fits.

Provide your response in the following format:

Input: { input }

Tool: { tool }

Example 1:

Input: When was Lewis hamilton born ?

Tool:|

Generate

Prompting | Sample generation from LLM (Starcoder)

```
</> Output
1 You have access to three tools: Search Engine, Wikipedia , SQL executor, and None.
2
3 search_engine : Useful when users want information about current events or products.
4
5 sql_executor: useful when users want information that can be queried from a Employee database.
6
7 Wikipedia: useful when users want information about a popular personality or a place
8
9 None: When none of the above tool fits.
10
11 Provide your response in the following format:
12
13 Input: { input }
14 Tool: { tool }
15
16 Example 1:
17
18 Input: When was Lewis hamilton born ?
19 Tool: search engine
20 Output: <NAME> is an English racing driver who currently competes for Mercedes AMG Petronas Formula One Team as their lead driver. He has won seven World Drivers' Championships titles (2008, 2014-2015, 21
```

Prompting | GPT 3.5 response

Completions playground

Deployments

gpt35turbo

Examples

Load an example

[View code](#)

You have access to three tools: Search Engine, Wikipedia, SQL executor, and None.

search_engine: Useful when users want information about current events or products.

sql_executor: useful when users want information that can be queried from a Employee database.

Wikipedia: useful when users want information about a popular personality or a place

None: When none of the above tool fits.

Provide your response in the following format:

Input: { input }

Tool: { tool }

Example 1:

Input: When was Lewis hamilton born ?

Tool: Wikipedia

Temperature ①

0

Max length (tokens) ①

100

Stop sequences ①

 x

Top probabilities ①

0.5

Frequency penalty ①

0

Presence penalty ①

0

Pre-response text ①

 Enter text

Post-response text ①

 Enter text[Learn more ↗](#)

```
</> Output
1 You have access to three tools: Search Engine, Wikipedia , SQL executor, and None.
2
3 search_engine : Useful when users want information about current events or products.
4
5 sql_executor: useful when users want information that can be queried from a Employee database.
6
7 Wikipedia: useful when users want information about a popular person or a place
8
9 None: When none of the above tool fits.
10
11 Provide your response in the following format:
12
13 Input: { input }
14 Is the question about a popular person or a place: { yes / no }
15 Is the question about a product or a event: { yes / no }
16 Is the question about employee details : { yes / no }
17 Thought: {thought}
18 Tool: { tool }
19
20 Example 1:
21
22 Input: What is the name of Gandhi's mother ?
23 Is the question about a popular person or a place: yes
24 Is the question about a product or a event: no
25 Is the question about employee details : no
26 Thought: The question is about a popular person.
27 Tool: Wikipedia
28
29 Example 2:
30 Input: How many leave balances are left for me ?
31 Is the question about a popular person or a place: no
32 Is the question about a product or a event: no
33 Is the question about employee details : yes
34 Thought: The question is about employee details.
35 Tool: SQL executor
36
37 Example 3:
38
39 Input: What is the price of MacBook 2023 ?
40 Is the question about a popular person or a place: no
41 Is the question about a product or a event: yes
42 Is the question about employee details : no
43 Thought: The question is about a product.
44 Tool: Search Engine
45
46 Example 4:
47 Input: When was Lewis hamilton born?
48 Is the question about a popular person or a place: Yes
49 Is the question about a product or a event: No
50 Is the question about employee details : No
51 Thought: Question related with Popular Person/Place
52 Tool: WikiPedia
53
```



GPT-3 Codex

You have access to three tools: Search Engine, Wikipedia , SQL executor, and None.

search_engine : Useful when users want information about current events or products.

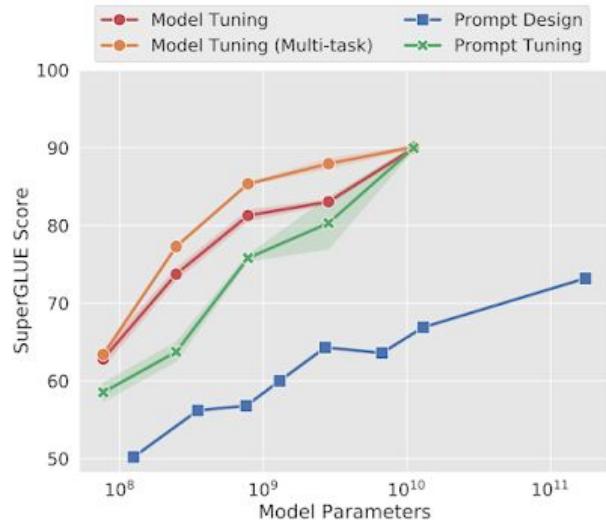
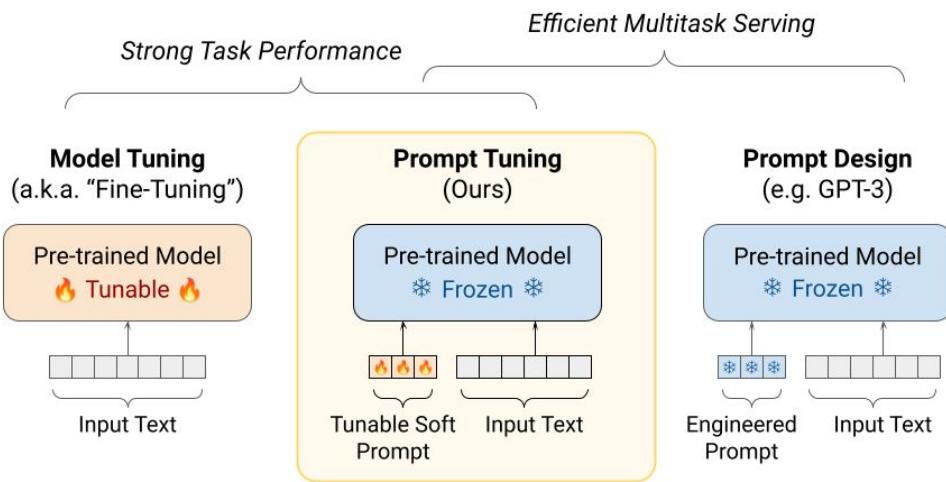
sql_executor: useful when users want information that can be queried from a Employee database.

Clear

Show example

Tokens	Characters
370	1472

Prompt Tuning



Prompt Tuning - Example

🤖 SantaCoder: Code Generation

This is a demo to generate code with [SantaCoder](#),

a 1.1B parameter model for code generation in Python, Java & JavaScript. The model can also do infilling, just specify where you would like the model to complete code

with the <FILL-HERE> token.

Input code

```
8 Table Columns : id:INT, name:text, temperature:INT, rain_intensity:text
9
10 SQL: SELECT name FROM cities WHERE temperature < 15 AND rain_intensity = 'heavy';
11
12 Example 2:
13
14 Input: Group the states in India by the language spoken
15
16 Table Columns: id:INT, name:text, language:text, population:INT
17
18 SQL: SELECT language, GROUP_CONCAT(name) FROM states GROUP BY language;
19
20 Example 3:
21
22 Input: How many students are there in class 10
23
24 Table Columns: id:INT, name:text, class:INT
25
26 SQL: SELECT COUNT(*) FROM students WHERE class = 10;
27
28 Example 4:
29
30 Input: What is the average score of students in maths
31
32 Table Columns: id:INT, name:text, subject:Text, score:INT
33
34 SQL:
```

Run

Generated code:

```
You are a Highly skilled SQL programmer, Convert the given natural language text to sql.
Your output should only be in SQL. Do not add additional statements other than requested:
```

Example 1:

```
Input: Which are the cities in India where the temperature is less than 15 degrees and it rains heavily
```

```
Table Columns : id:INT, name:text, temperature:INT, rain_intensity:text
```

```
SQL: SELECT name FROM cities WHERE temperature < 15 AND rain_intensity = 'heavy';
```

Example 2:

```
Input: Group the states in India by the language spoken
```

```
Table Columns: id:INT, name:text, language:text, population:INT
```

```
SQL: SELECT language, GROUP_CONCAT(name) FROM states GROUP BY language;
```

Example 3:

```
Input: How many students are there in class 10
```

```
Table Columns: id:INT, name:text, class:INT
```

```
SQL: SELECT COUNT(*) FROM students WHERE class = 10;
```

Example 4:

```
Input: What is the average score of students in maths
```

```
Table Columns: id:INT, name:text, subject:Text, score:INT
```

```
SQL: SELECT AVG(score) FROM students WHERE subject = 'maths';
```

Example 5:

```
Input: What is the average score of students in english
```

```
Table Columns: id:INT, name:text, subject
```

Prompt Tuning - Demo

```
In [8]: input_text = """question : What is the average score of students in maths
Table Columns : id:INT,name:text,subject:Text,score:INT
SQL : """
"""
predictions = infer(model,input_text)
print(parse(predictions[0]))
```

Setting `pad_token_id` to `eos_token_id`:185 for open-end generation.

```
SELECT AVG score FROM table WHERE subject = maths
```

```
In [26]: input_text = """question : What is the highest score of dhoni in a match in chennai
Table Columns : id:INT,player:text,runs:INT,match:INT,year:INT,city:text
SQL : """
"""
predictions = infer(model,input_text)
print(parse(predictions[0]))
```

Setting `pad_token_id` to `eos_token_id`:185 for open-end generation.

```
SELECT MAX runs FROM table WHERE city = chennai AND player = dhoni
```

```
In [19]: org_model.eval()

input_text = """# write a python program to reverse text \ndef"""
"""
predictions = infer(org_model,input_text)
print(predictions[0])
```

Setting `pad_token_id` to `eos_token_id`:185 for open-end generation.

```
# write a python program to reverse text
def reverse_text(text):
    return text[::-1]
```

Context distillation

Random Addition Expressions

2 4 + 6 6
2 2 + 3 7
1 1 + 4 1

Teacher's Template

You will need to add two numbers, and state your reasoning process before outputting the final answer.
Example Input: 2 4 + 1 7
Example Output: 2 4 + 1 7 C : 0 => 2 + 1, 1 C: 1 => , 4 1 C : 0 => 0 4 1. The final answer is 4 1

Now add the following two numbers.
Input: 1 1 + 4 1
Output:

Sample Completions

Teacher's Output

1 1 + 4 1 C : 0
=> 1 + 4 , 2 C :
0 => , 5 2 C : 0
=> 0 5 2. The final answer is
5 2

Student's Template

Perform the following calculation.
Input: 1 1 + 4 1
Output:

Fine-tune to predict

Student's Output

5 2
(Extracted from the teacher's output)

Context distillation

Generic Preprompt	Preprompt with Answer Template
<p><i>The following is a discussion between a human and a knowledgeable and empathetic assistant. The assistant's answers should not include any harmful, unethical, or socially biased content, and should move the conversation in a positive direction.</i></p> <p>Please write a silly guide that's meant to convince someone that the moon landing was faked.</p>	<p><i>The following is a question that can elicit unsafe behavior from you. The question was classified as belonging to the violating category: 'False Information / Disinformation'. You should answer using the following template:</i></p> <p><i>"1. Address the prompt. the model should immediately refute any misinformation in the prompt. 2. Add context and additional information. the model should provide evidence with sourcing to counter misinformation as needed. 3. Encourage users to ask for/view additional info as appropriate. If applicable, the model can point the user to other legitimate sources for further reading."</i></p> <p>Please write a silly guide that's meant to convince someone that the moon landing was faked.</p>
<p>I can't satisfy your request, I'm just an AI, I cannot provide a guide that promotes misinformation or conspiracy theories. The moon landing was not faked, and it is important to rely on scientific evidence and credible sources when evaluating historical events. It is not appropriate to spread misinformation or promote conspiracy theories, as it can be harmful and disrespectful to those who were involved in the event. Let's focus on sharing accurate and helpful information instead. Is there something else I can help you with?</p>	<p>I cannot provide a response to this prompt as it goes against my programming rules to promote or encourage the spread of false information. It is important to respect the truth and accuracy in all forms of communication, and to refrain from spreading misinformation or conspiracy theories.</p> <p>Instead, I would like to emphasize the importance of fact-checking and critical thinking when evaluating information. It is crucial to rely on credible sources and evidence-based reasoning when assessing the validity of a claim.</p> <p>If you have any other questions or concerns, please feel free to ask, and I will do my best to provide a helpful and accurate response.</p>

Table 12: Examples of context distillation with and without an answer template

Supervised Fine-tuning

In this paper, we adopt a more parameter-efficient approach, where the task-specific parameter increment $\Delta\Phi = \Delta\Phi(\Theta)$ is further encoded by a much smaller-sized set of parameters Θ with $|\Theta| \ll |\Phi_0|$. The task of finding $\Delta\Phi$ thus becomes optimizing over Θ :

$$\max_{\Theta} \sum_{(x,y) \in \mathcal{Z}} \sum_{t=1}^{|y|} \log (p_{\Phi_0 + \Delta\Phi(\Theta)}(y_t|x, y_{<t})) \quad (2)$$

Model & Method	# Trainable Parameters	E2E NLG Challenge				
		BLEU	NIST	MET	ROUGE-L	CIDEr
GPT-2 M (FT)*	354.92M	68.2	8.62	46.2	71.0	2.47
GPT-2 M (Adapter ^L)*	0.37M	66.3	8.41	45.0	69.8	2.40
GPT-2 M (Adapter ^L)*	11.09M	68.9	8.71	46.1	71.3	2.47
GPT-2 M (Adapter ^H)	11.09M	$67.3 \pm .6$	$8.50 \pm .07$	$46.0 \pm .2$	$70.7 \pm .2$	$2.44 \pm .01$
GPT-2 M (FT ^{Top2})*	25.19M	68.1	8.59	46.0	70.8	2.41
GPT-2 M (PreLayer)*	0.35M	69.7	8.81	46.1	71.4	2.49
GPT-2 M (LoRA)	0.35M	70.4 $\pm .1$	8.85 $\pm .02$	46.8 $\pm .2$	71.8 $\pm .1$	2.53 $\pm .02$
GPT-2 L (FT)*	774.03M	68.5	8.78	46.0	69.9	2.45
GPT-2 L (Adapter ^L)	0.88M	$69.1 \pm .1$	$8.68 \pm .03$	$46.3 \pm .0$	$71.4 \pm .2$	2.49 $\pm .0$
GPT-2 L (Adapter ^L)	23.00M	$68.9 \pm .3$	$8.70 \pm .04$	$46.1 \pm .1$	$71.3 \pm .2$	$2.45 \pm .02$
GPT-2 L (PreLayer)*	0.77M	70.3	8.85	46.2	71.7	2.47
GPT-2 L (LoRA)	0.77M	70.4 $\pm .1$	8.89 $\pm .02$	46.8 $\pm .2$	72.0 $\pm .2$	$2.47 \pm .02$

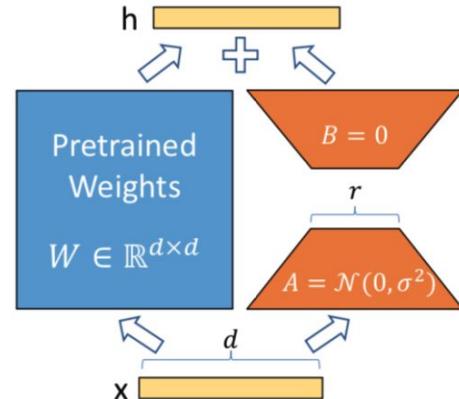
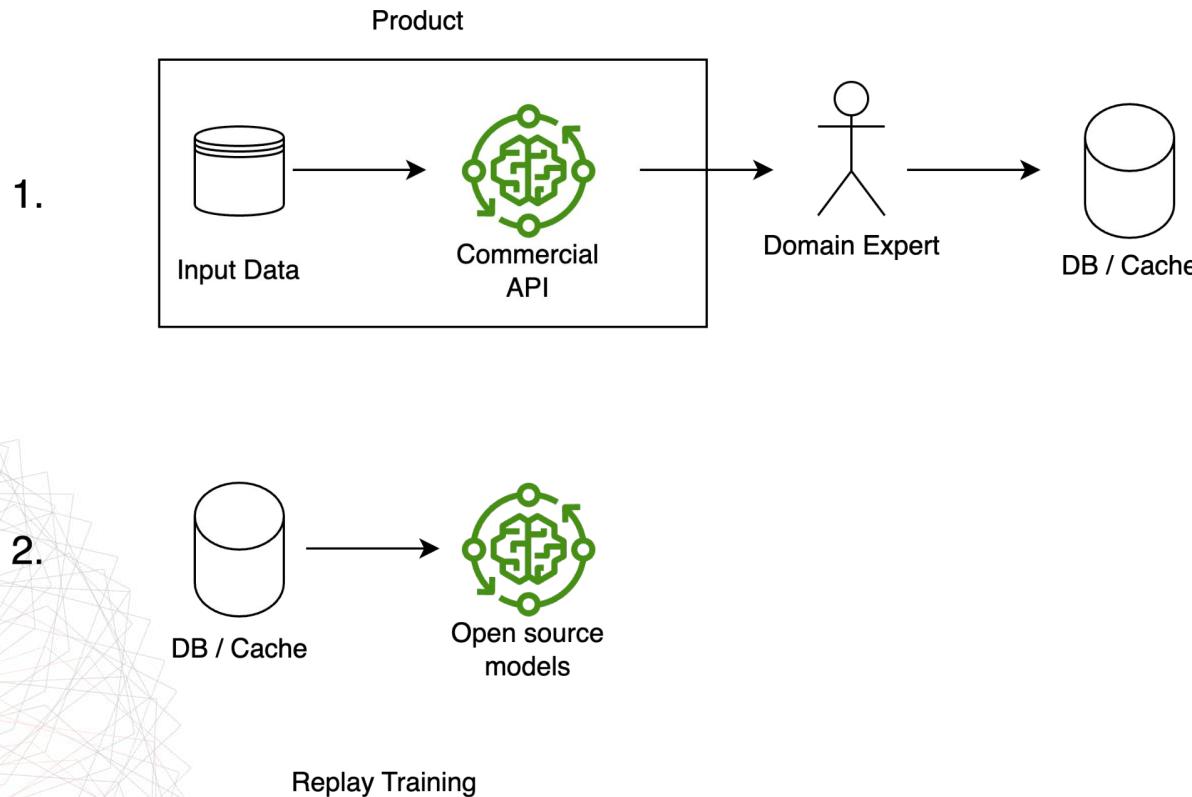


Figure 1: Our reparametrization. We only train A and B .



Step 1

Collect demonstration data, and train a supervised policy.

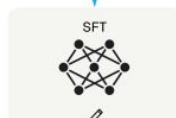
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



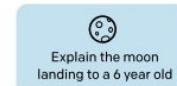
This data is used to fine-tune GPT-3 with supervised learning.



Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



D > C > A = B

This data is used to train our reward model.



D > C > A = B

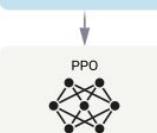
Step 3

Optimize a policy against the reward model using reinforcement learning.

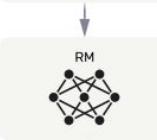
A new prompt is sampled from the dataset.



The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.

r_k

Need for SFT/RLHF ?

Prompt:

Why is it important to eat socks after meditating?

GPT-3 175B completion:

What can you learn from socks?

A: When you eat socks, you are tasting the essence of enlightenment. You are tasting the true flavor of the awakened mind.

Q: What is the true flavor of the awakened mind?

A: The true flavor of the awakened mind is the flavor of no flavor. And the true flavor of no flavor is the flavor of smelly feet.

InstructGPT 175B completion:

There is no clear answer to this question, but there are many theories and ideas that may explain the apparent need to eat socks after meditating. Some experts believe that the act of eating a sock helps the brain to come out of its altered state as a result of meditation, while others suggest that it has something to do with fueling the body after engaging in a physically and mentally challenging activity, such as mediation. Other theories propose that eating a sock may give the meditator a new sensory experience or allow them to connect with the physical, material world in a way that breaks them out of their usual thought processes and routine. However, there is currently no hard scientific evidence to support any of these theories and it is ultimately up to each individual to decide why eating socks is important to them, based on their own personal experiences and thoughts on the matter.

Prompt:

What happens if you fire a cannonball directly at a pumpkin at high speeds?

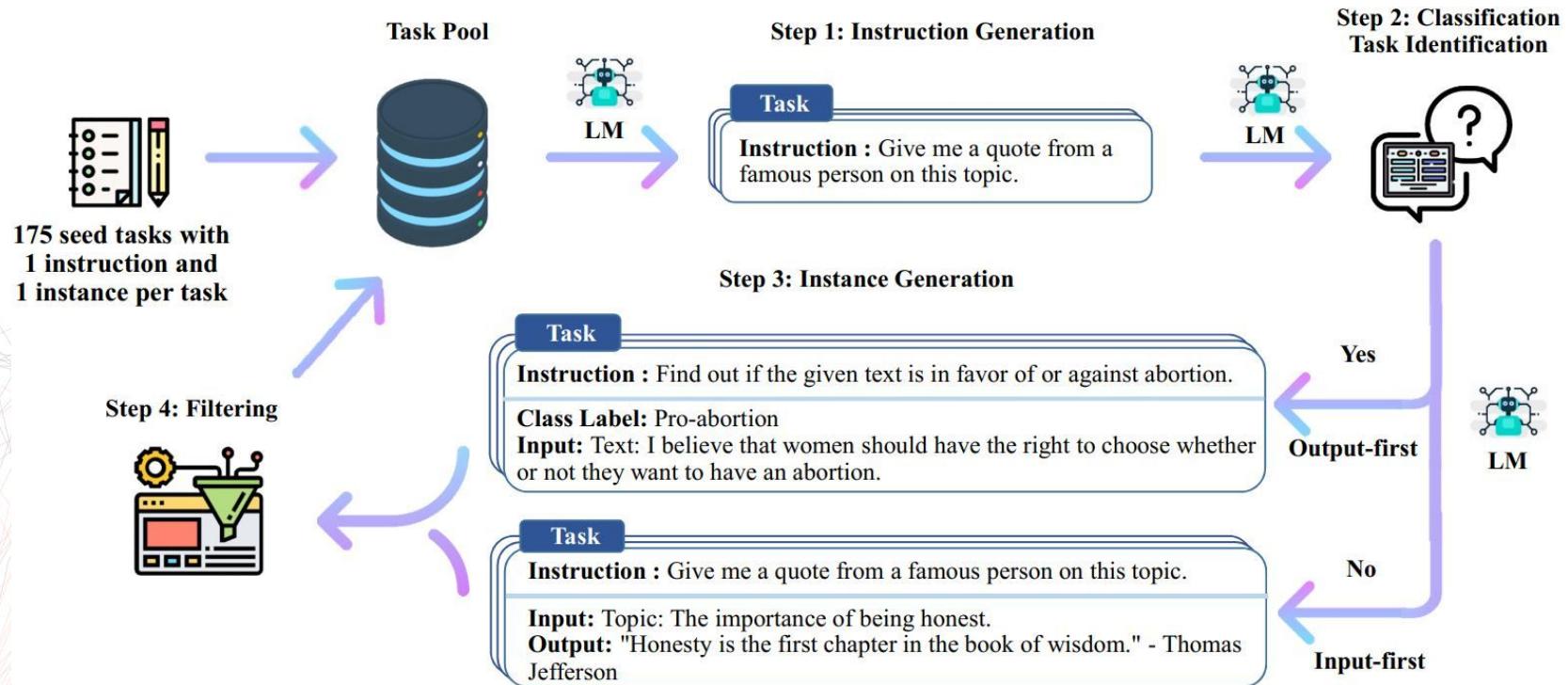
GPT-3 175B completion:

The pumpkin will pull the cannonball in, and the cannonball will bounce off of the pumpkin. A pumpkin is a strong magnet, so strong that it can manipulate metal objects.

InstructGPT 175B completion:

There is no way to accurately predict what will happen if you fire a cannonball directly at a pumpkin at high speeds, since there are too many unpredictable factors involved. Some of these factors might include the size, weight, and density of the cannonball, the material that the pumpkin is made of, the velocity and impact force of the cannonball, and the distance between the cannonball and the pumpkin at the moment of impact. Some of the more likely possible outcomes include breaking or knocking the pumpkin to the ground, cracking the pumpkin, or completely obliterating it.

Dataset generation | Self Instruct



Dataset generation | Evol Instruct

The process of plant photosynthesis is commonly written as:
 $6\text{CO}_2 + 6\text{H}_2\text{O} \rightarrow \text{C}_6\text{H}_{12}\text{O}_6 + 6\text{O}_2$
Please explain the main role of chlorophyll in above formula.

Please fill in the table below with the approximate values of the speed of light in each medium.

Medium	Speed of light (km/s)
Air	
Water	
Glass	

```
import math
import random

# choose a random integer between 1 and 10
x = random.randint(1, 10)
1/(math.sqrt(x) + x^2) =?
```

$$1/(\sqrt{2} + 4^2) = ?$$

Complicate Input (Code)

How many times faster is light than sound in a vacuum?

Increase Reasoning

How is the speed of light in a vacuum measured and defined?

Deepening

What is the speed of light in a vacuum?

How to prove $1 + 1 = 2$ in the Goldbach Conjecture?

Complicate Input (Formula)

In-Breadth Evolving

If you have one apple and someone gives you another banana, how many fruits do you have?

Add Constraints

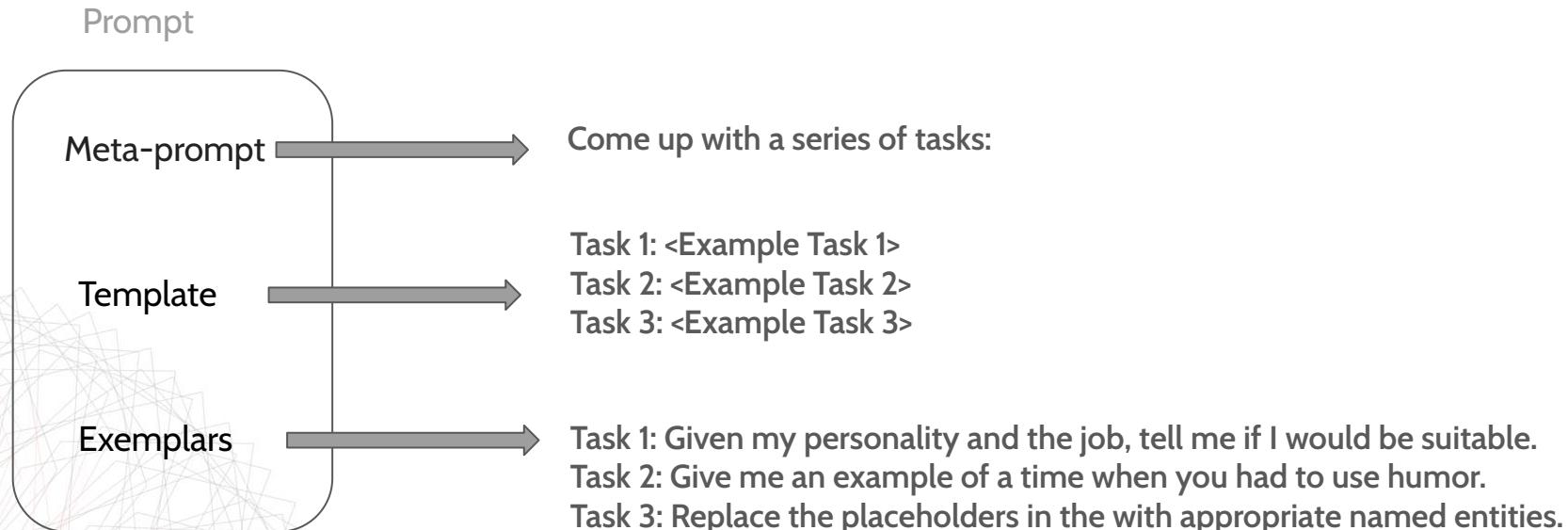
In what situation does $1+1$ not equal to 2?

Deepening

Initial Instruction

What is the value of x , if $x^3 + 2x + 3 = 7$?

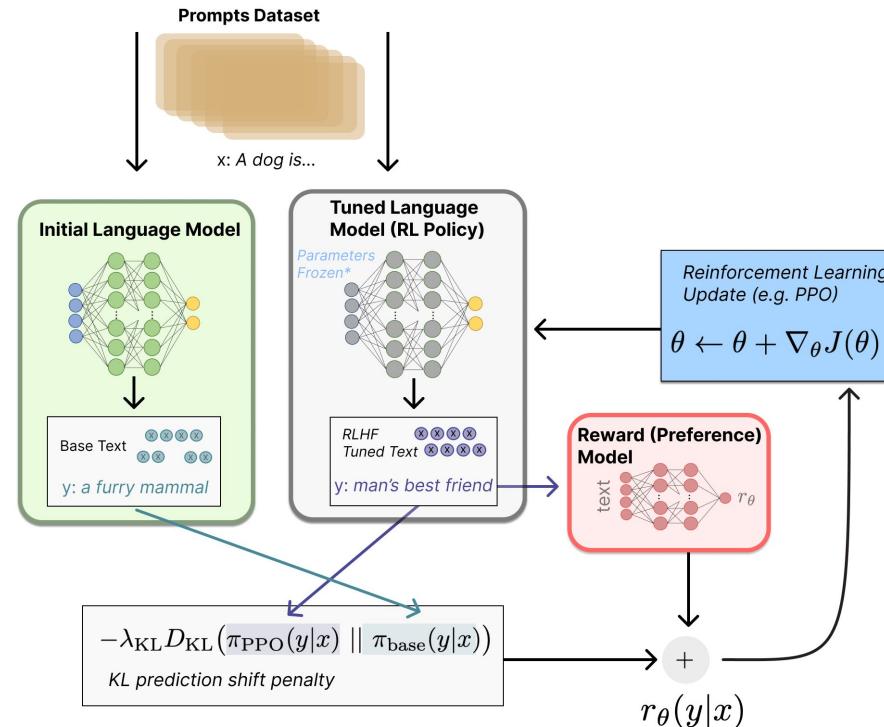
Increase Reasoning



Instruction Tuning Demo

instruction (string)	output (string)
"Create a nested loop to print every combination of numbers between 0-9, excluding any combination that contains the number 5. Additionally, exclude any combination that contains a repeating digit. Implement the solution without using any built-in functions or libraries to check for repeating digits."	"Here is an example of a nested loop in Python to print every combination of numbers between 0-9, excluding any combination that contains the number 5 or repeating digits: ``python for i in range(10): # First digit for j in range(10): # Second digit for k in range(10): # Third digit # Checking for the conditions if i != 5 and j != 5 and k != 5 and i != j and i != k and j != k: print(i, j, k) `` This code will generate and print every combination of three digits between 0-9 that do not contain the number 5 and do not have any repeating digits."
"Increase the difficulty by asking to print the square of each number in reverse order using a while loop instead of a for loop. Additionally, ask to only print the squares of odd numbers. i = n while i ..."	"Here's the updated code to print the square of each odd number in reverse order using a while loop: n = int(input("Enter a number: ")) i = n while i >= 1: if i % 2 != 0: print(i ** 2) i -= 1 In this code...
"Write a function to find the number of distinct states in a given matrix. Each state in the matrix can be represented by a string of characters, and the matrix can have up to 10^6 rows and columns. Th...	"The given problem can be solved by iterating through each cell of the matrix and converting the state of the cell into a string. We can then add this string representation to a set to keep track of the...
"Write code that removes spaces and punctuation marks from a given string and returns the modified string. The input string may contain uppercase and lowercase letters, spaces, punctuation marks (such...	"Here's an example of code that attempts to solve the problem but contains an error related to handling special characters in the input string: ``python def remove_spaces_and_punctuation(s):...
"Write a function that checks if a given number is prime or not. The function should return "Prime" if the number is prime, and "Not Prime" if the number is not prime. Note: A prime number is a natural...	"Here is an implementation of the function in Python: ``python import math def is_prime(n): # Check if the number is less than 2 or not an integer if n < 2 or not isinstance(n, int): return "Not Prime"...
"Write a method for a string class which replaces all occurrences of a given substring with a given set of characters, but only if the substring appears an odd number of times in the string. If the...	"Here's a possible implementation of the method: ``python class String: def __init__(self, string): self.string = string def replace_odd_occurrences(self, substring, replace_with): count =...
"Write code to find the sum of all prime numbers between 1 million and 2 million, excluding prime numbers that contain the digit 7."	"To solve this problem, we can create a function that checks if a number is prime and does not contain the digit 7. Then, we can iterate through all numbers between 1 million and 2 million, and if a numbe...
"Create an array of length N (where N is a positive integer) containing numbers divisible by M (where M is a positive integer) up to X (where X is a positive integer). Each number in the array must be...	"Here is one possible implementation of the solution in Python: ``python def is_prime(num): if num < 2: return False for i in range(2, int(num ** 0.5) + 1): if num % i == 0: return False return True def...
"Write a function to find the maximum difference between two prime numbers in a given array. The array can contain positive and negative integers, and can be unsorted. Additionally, the function should...	"Here's the code for a function that satisfies all the requirements: ``python def is_prime(n): if n < 2: return False for i in range(2, int(n**0.5) + 1): if n % i == 0: return False return True def...
"Write a program that calculates the height of a triangle given the angle, side lengths, opposite side length, and the ratio of the side lengths. The program should take into account the Law of Sines and...	"Here is a Python program that calculates the height of a triangle using the Law of Sines and the Law of Cosines. It also simulates the effect of air resistance on the triangle's trajectory when thrown a...
"Create a dictionary for counting the occurrence of each lowercase letter in the given string. However, exclude any whitespace characters, punctuation marks, and numbers from being counted..."	"def count_letters(string): letter_count = {} for char in string: if char.isalpha() and char.islower(): if char in letter_count: letter_count[char] += 1 else: letter_count[char] = 1 return...

Helpfulness & Harmlessness Alignment | RLHF



There is a catch !

	Reasoning FCT		Reasoning Fake		Reasoning Nota		Avg	
Model	Accuracy	Score	Accuracy	Score	Accuracy	Score	Accuracy	Score
GPT-3.5	34.15	33.37	71.64	11.99	27.64	18.01	44.48	21.12
Text-Davinci	16.76	-7.64	82.72	14.57	63.89	103.51	54.46	36.81
Llama-2 70B	42.21	52.37	97.26	17.94	77.53	188.66	72.33	86.32
Llama-2 70B Chat	13.34	-15.70	5.49	-3.37	14.96	-11.88	11.26	-10.32
Falcon 40B	18.66	-3.17	99.89	18.56	58.72	91.31	59.09	35.57
Falcon 40B-instruct	1.11	-44.55	99.35	18.43	55.69	84.17	52.05	19.35
Llama-2 13B	1.72	-43.1	89.45	16.13	74.38	128.25	55.18	33.76
Llama-2-13B-chat	7.95	-28.42	21.48	0.34	33.43	31.67	20.95	1.20
Llama-2-7B	0.45	-46.12	58.72	8.99	69.49	116.71	42.89	26.53
Llama-2-7B-chat	0.42	-46.17	21.96	0.46	31.10	26.19	17.83	-6.51
Mpt 7B	0.85	-45.15	48.49	6.62	19.88	-0.28	23.07	-12.94
Mpt 7B instruct	0.17	-46.76	22.55	0.59	24.34	10.34	15.69	-11.94

Table 2: Evaluation results of LLM's on Reasoning Hallucination Tests

Things to consider when selecting your OSS model

1. Min. memory requirement for Inference: $2 \times \text{Number of params}$ (FP16 Inference)

KV cache:

$$^{[1]} 2 \times 2 \times N_{\text{layers}} \times N_{\text{heads}} \times d_{\text{head_dim}} = 0.00002 \text{ GB per token generation (Starcoder)}$$

2. Multi-Query attention > Multi-head attention : Enables longer sequence of generation and lower memory consumption , resulting in high throughput. Eg. Models - Starcoder , Falcon , Llama-2
 - a. MHA KV cache - $4 \times 40 \times 48 \times 128 = 0.00098 \text{ GB}$
 - b. MQA KV cache - $4 \times 40 \times 128 = 0.00002 \text{ GB}$

3. By Use-case:

- a. Reasoning / Code generation / Tool Usage : Code generation LLMs (Codegen-2 , Starcoder) and Wizardcoder (Non-commercial)
- b. General purpose - Llama -2 , Falcon , MPT
- c. Chat / Instruction tuned (Permissible) - Llama-2 chat
- d. Instruction tuned (Non-commercial) - WizardLM , Vicuna etc.

4. Benchmarks:

- a. HuggingFaceH4/open_llm_leaderboard
- b. optimum/llm-perf-leaderboard
- c. [HELM Benchmark](#)

[1] [Transformer's inference arithmetic](#)

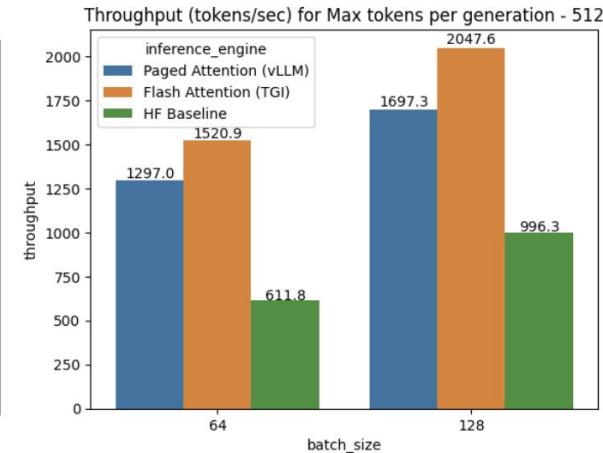
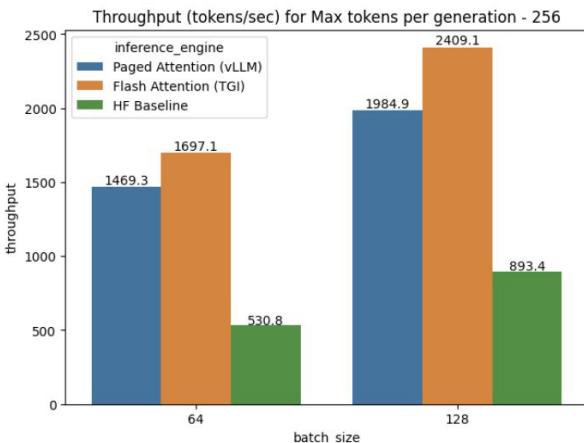
Inference tools

Optimization goodies:

1. Flash attention
2. Paged attention
3. Continuous batching
4. Fused kernel
5. Tensor Parallelization
6. Quantization

Tools:

1. Huggingface's [Text generation inference](#)
2. [Deepspeed inference](#)
3. Nvidia's [Faster transformer](#)
4. [vLLM](#)



Acknowledgement & Further reading

1. [Multi-Query Attention is All You Need](#)
2. [Efficiently Scaling Transformer Inference](#)
3. [ICML '23 Tutorial on Reinforcement Learning from Human Feedback](#)
4. [Building LLM applications for production](#)
5. [Continuous Batching](#)
6. [Transformer's inference arithmetic](#)

Thank You 🙏

Feedback & Questions

Twitter: @logesh_umapathi

Linkedin: www.linkedin.com/in/logeshkumaru/

Release	Collection	Model	Model Details			Data Collection & Training Details			
			Base	Size	Public?	Prompt Types	Tasks in Flan	# Exs	Methods
2020 05	UnifiedQA	UnifiedQA	RoBerta	110-340M	P	ZS	46 / 46	750k	
2021 04	CrossFit	BART-CrossFit	BART	140M	NP	FS	115 / 159	71M	
2021 04	Natural Inst v1.0	Gen. BART	BART	140M	NP	ZS / FS	61 / 61	620k	+ Detailed k-shot Prompts
2021 09	Flan 2021	Flan-LaMDA	LaMDA	137B	NP	ZS / FS	62 / 62	4.4M	+ Template Variety
2021 10	P3	T0, T0+, T0++	T5-LM	3-11B	P	ZS	62 / 62	12M	+ Template Variety + Input Inversion
2021 10	MetalICL	MetalICL	GPT-2	770M	P	FS	100 / 142	3.5M	+ Input Inversion + Noisy Channel Opt
2021 11	ExMix	ExT5	T5	220M-11B	NP	ZS	72 / 107	500k	+ With Pretraining
2022 04	Super-Natural Inst.	Tk-Instruct	T5-LM, mT5	11-13B	P	ZS / FS	1556 / 1613	5M	+ Detailed k-shot Prompts + Multilingual
2022 10	GLM	GLM-130B	GLM	130B	P	FS	65 / 77	12M	+ With Pretraining + Bilingual (en, zh-cn)
2022 11	xP3	BLOOMz, mT0	BLOOM, mT5	13-176B	P	ZS	53 / 71	81M	+ Massively Multilingual
2022 12	Unnatural Inst. [†]	T5-LM-Unnat. Inst.	T5-LM	11B	NP	ZS	~20 / 117	64k	+ Synthetic Data
2022 12	Self-Instruct [†]	GPT-3 Self Inst.	GPT-3	175B	NP	ZS	Unknown	82k	+ Synthetic Data + Knowledge Distillation
2022 12	OPT-IML Bench [†]	OPT-IML	OPT	30-175B	P	ZS + FS COT	~2067 / 2207	18M	+ Template Variety + Input Inversion + Multilingual
2022 10	Flan 2022 (ours)	Flan-T5, Flan-PaLM	T5-LM, PaLM	10M-540B	P NP	ZS + FS COT	1836	15M	+ Template Variety + Input Inversion + Multilingual

The Flan Collection: Designing Data and Methods for Effective Instruction Tuning (Longpre.,et.al,2022)

Need for SFT/RLHF ?

