

I used Gemma-2b-it and Llama-3.1-8B-Instruct from together API and Phi-3.5-mini-instruct from HuggingFace.

Accuracies:

```
Gemma Zero-shot accuracy: 28.00%  
LLaMA Zero-shot accuracy: 30.00%  
Gemma Chain of Thought accuracy: 24.00%  
LLaMA Chain of Thought accuracy: 26.00%
```

reAct gemma: 0.25

reAct llama: 0.37

Phi Accuracies:

Zero-shot accuracy: 0.28

Chain-of-Thought accuracy: 0.28

ReAct accuracy: 0.34

Inference times per question:

Gemma Zero Shot time = approx 0.5 to 2 sec

Gemma CoT = approx 1 to 2 sec

Gemma reAct = approx 3 to 5 sec

LlMa Zero Shot time = approx 1 sec

LlMa CoT = approx 2 sec

LlMa reAct = approx 2 to 3 sec

Phi Zero Shot time = approx 0 to 2 sec

Phi CoT = approx 0 to 3 sec

Phi reAct = approx 5 to 7 sec

The inference time of reAct is much higher, it uses an external tool LLMSymbolicMathChain with a bigger thought, reasoning, action prompting chain. CoT is higher than zero shot because it produces the sequence of steps the model takes to solve the problem.

The output quality is generally much better for reAct prompting due to detailed stepwise answers and use of external tools for calculations, chain of thought had the worst accuracy, probably due to hallucinations. Zero shot outputs answers directly in one line mostly.

The best performing model was LlMa, it has 8B parameters and is a better model.

Bigger models with more parameters take more time to load and perform inference, have a lot of hardware requirements of GPU and CPU.

Could not compare time with Phi since it was not available on Together.ai had to use it directly from huggingface which took more time, but it should have lower inference time due to optimisations in the model and lower parameters (3.8B).

Gemma is just 2B parameter so it is the fastest of the three models and runs even on CPU.

LlMa has 8B parameters so takes longer time, but since it has better performance, it trades speed for accuracy and is quite powerful.

Phi struggles with reasoning required tasks but achieves good performance for its size. It struggles with multi-step reasoning however and has the middle accuracy out of the three. Gemma is also a lightweight model, and it performs better for text generation and english language prompts but is unable to handle multi step reasoning tasks, hence performing the worst out of the three models.

Zero-shot prompting gives quick responses but can be less accurate with more complex questions. It works well with smaller models like Gemma-2B for straightforward queries. Chain of Thought (zero-shot) helps models reason through tasks step by step, improving accuracy, particularly for Llama-3.1 and Phi-3.5-mini, but it increases inference time as models perform deeper reasoning.

ReAct Prompting introduces more complex reasoning and action-taking steps, and while it boosts accuracy across all models, it is computationally expensive, especially for large models like Meta-Llama-3.1.

Trade Offs: Model Size Accuracy vs. Speed and Prompting techniques output quality vs inference time

Example:

question = "What is the integral of x^2 ?"

options = "a) $x^3/3$, b) $x^2/2$, c) $2x$, d) $1/x$ "

Gemma 2B IT Zero-shot: Correct answer: a) $x^3/3$

Explanation: The integral of x^2 is $x^3/3$., Time: 0.6204948425292969

Gemma 2B IT Chain of Thought: The integral of x^2 is $x^3/3$.

Step 1: Integrate x^2 with respect to x .

$\int x^2 dx = x^3/3 + C$, where C is the constant of integration.

The answer is $x^3/3 + C$., Time: 1.0094835758209229

LLaMA 8B Zero-shot: The correct answer is a) $x^3/3$.

To find the integral of x^2 , we use the power rule of integration, which states that:

$\int x^n dx = (x^{(n+1)})/(n+1) + C$

where n is a constant, and C is the constant of integration.

For x^2 , $n = 2$, so:

$\int x^2 dx = (x^{(2+1)})/(2+1) + C$

$= (x^3)/3 + C$

Therefore, the integral of x^2 is $x^3/3$., Time: 2.0097410678863525

LLaMA 8B Chain of Thought: To find the integral of x^2 , we need to follow the power rule of integration, which states that the integral of x^n is $(x^{(n+1)})/(n+1)$.

In this case, $n = 2$, so we substitute n into the formula:

$\int x^2 dx = \int x^{(2+1)} dx$

$= \int x^3 dx$

$= (x^{(3+1)})/(3+1)$

$$= x^4/4$$

However, none of the given options match this result. But if we consider the option $x^2/2$, we can see that it's close to the actual result. However, the actual result is $x^4/4$, which is not among the given options.

However, we can see that if we take the derivative of $x^2/2$, we get x , so we can say that $x^2/2$ is the antiderivative of x , but the question asks for the integral of x^2 , which is actually the antiderivative of x^2 . Time: 1.2032525539398193

Prompt: reAct Model: GEMMA

Generated text:

****Answer: 1. $x^3/3$ ****

****Explanation:****

The integral of x^2 is $(x^3)/3$. This is a straightforward integral that can be solved by using the power rule of integration.

Extracted answer: 1

Inference time: 0.8958 seconds

Prompt: reAct Model: LLaMa

Generated text:

Let's break down the problem step by step.

1. First, identify the type of mathematical problem we are dealing with.

- The problem is asking for the integral of x^2 . In mathematical terms, we need to find the antiderivative or the indefinite integral of the function $f(x) = x^2$.

2. Analyze the given choices.

- Let's go through each choice and consider if it could be a potential solution:

1. $x^3/3$

2. $x^2/2$

3. $2x$

4. $1/x$

3. Perform necessary calculations or manipulations to arrive at the correct solution.

- To find the integral of x^2 , we can use the power rule of integration, which states that $\int x^n dx = (x^{n+1})/(n+1) + C$, where n is a constant and C is the constant of integration.

- In this case, $n = 2$, so we have:

$$\int x^2 dx = (x^{2+1})/(2+1) + C$$

$$\int x^2 dx = (x^3)/3 + C$$

4. Select the correct choice based on the calculations.

- Based on the calculations, the correct solution is:

$$\int x^2 dx = (x^3)/3 + C$$

Therefore, the correct answer is:

Final Answer: 1

Explanation: The correct solution is $x^3/3$, which matches choice 1. This is because the integral of x^2 is indeed $(x^3)/3$, as derived using the power rule of integration.

Extracted answer: 1

Inference time: 1.4801 seconds

Phi-3.5 Zero-Shot

The correct answer to the question "What is the integral of x^2 ?" is "a) $x^3/3$ ", Time: 0.3243 seconds

Phi-3.5 Chain of Thought

To find the integral of x^2 , we will apply the power rule for integration. The power rule states that the integral of $x^n dx$ is equal to $(x^{(n+1)})/(n+1) + C$, where n is a constant and C is the integration constant.

Here, $n = 2$. So, we will apply the power rule as follows:

$$\text{Integral of } x^2 dx = (x^{(2+1)})/(2+1) + C = (x^3)/3 + C$$

Comparing the result to the given options, we find that the correct answer is:

a) $x^3/3$ Yes, the correct answer is option a) $x^3/3$., Time: 1.5324 seconds

Phi-3.5 reAct

Final Answer: 1. $x^3/3$ Step 1: Identify the type of mathematical problem.

The given problem involves finding the integral of a function. Therefore, this is an integration problem.

Step 2: Analyze the given choices.

The choices provided are potential answers for the integral of x^2 . We need to choose the correct one.

Step 3: Perform necessary calculations using the math chain.

To find the integral of x^2 , we will use the power rule: $\int x^n dx = (x^{(n+1)})/(n+1) + C$, where C is the constant of integration. In our case, $n=2$. So, $\int x^2 dx = (x^{(2+1)})/(2+1) + C = x^3/3 + C$

Step 4: Based on the calculations, select the correct choice.

The correct answer is 1. " $x^3/3$ " since it matches our calculated value. Final Answer: 1. " $x^3/3$ " Given the question, you are asked to find the integral of x^2 . We know that the integral of x^n is $(x^{(n+1)})/((n+1)) + C$, where C is the constant of integration.

Applying this formula to x^2 , we get:

$$\begin{aligned}\int x^2 dx &= (x^{(2+1)})/((2+1)) + C \\ &= x^3/3 + C\end{aligned}$$

Now let's compare this to our given choices:

$x^3/3$ (matches our calculated value)

$x^2/2$ (incorrect, derived from integrating x)

$2x$ (incorrect, derived from integrating a constant function)

$1/x$ (incorrect, derived from integrating a negative power function)

In conclusion, the correct choice is 1. " $x^3/3$ ". Final Answer: 1. " $x^3/3$ " The given problem involves finding the indefinite integral of x^2 , which is a fundamental concept in calculus known as finding antiderivatives or primitive functions., Time: 12.4830 seconds

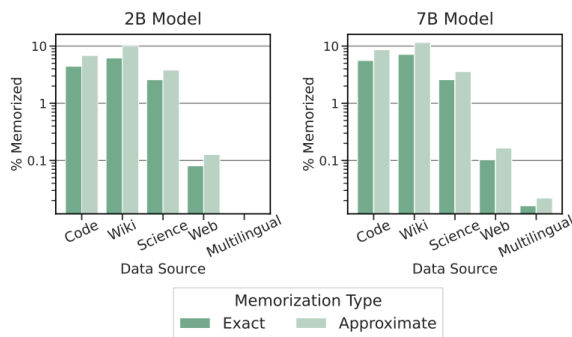
Introduction:

According to Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone, Phi-3.5 archives an accuracy of 68.8% whereas LLaMa 8B achieved 66.0% (5-shot prompting). LLaMa 8B has more number of number of parameters, but Phi 3.5 performs well for its size, because of its compactness and efficiency. It uses high-quality data and advanced optimizations, such as quantization, but it still behind Meta-Llama due to its smaller parameter count in zero shot and chain of thought prompting. Phi “simply does not have the capacity to store too much “factual knowledge”.” However, it is better than Gemma in maths due to its emphasis on reasoning tasks during training. Gemma is smallest model (2 billion parameters) out of the three, limiting its capacity for handling complex and multi-step tasks like college-level mathematics. It is designed more for practical use, focusing on speed and efficiency, which trades for reasoning abilities.

Training Datasets Composition:

Category	Benchmark	Llama 3 8B	Gemma 2 9B	Mistral 7B	Llama 3 70B	Mixtral 8x22B	GPT 3.5 Turbo	Llama 3 405B	Nemotron 4 340B	GPT-4o (0125)	GPT-4o	Claude 3.5 Sonnet
General	MMLU (5-shot)	69.4	72.3	61.1	83.6	76.9	70.7	87.3	82.6	85.1	89.1	89.9
	MMLU (0-shot, CoT)	73.0	72.3 [△]	60.5	86.0	79.9	69.8	88.6	78.7 [△]	85.4	88.7	88.3
	MMLU-Pro (5-shot, CoT)	48.3	–	36.9	66.4	56.3	49.2	73.3	62.7	64.8	74.0	77.0
	IFEval	80.4	73.6	57.6	87.5	72.7	69.9	88.6	85.1	84.3	85.6	88.0
Code	HumanEval (0-shot)	72.6	54.3	40.2	80.5	75.6	68.0	89.0	73.2	86.6	90.2	92.0
	MBPP EvalPlus (0-shot)	72.8	71.7	49.5	86.0	78.6	82.0	88.6	72.8	83.6	87.8	90.5
Math	GSM8K (8-shot, CoT)	84.5	76.7	53.2	95.1	88.2	81.6	96.8	92.3 [◇]	94.2	96.1	96.4 [◇]
	MATH (0-shot, CoT)	51.9	44.3	13.0	68.0	54.1	43.1	73.8	41.1	64.5	76.6	71.1
Reasoning	ARC Challenge (0-shot)	83.4	87.6	74.2	94.8	88.7	83.7	96.9	94.6	96.4	96.7	96.7
	GPQA (0-shot, CoT)	32.8	–	28.8	46.7	33.3	30.8	51.1	–	41.4	53.6	59.4
Tool use	BFCL	76.1	–	60.4	84.8	–	85.9	88.5	86.5	88.3	80.5	90.2
	Nexus	38.5	30.0	24.7	56.7	48.5	37.2	58.7	–	50.3	56.1	45.7
Long context	ZeroSCROLLS/QuALITY	81.0	–	–	90.5	–	–	95.2	–	95.2	90.5	90.5
	InfiniteBench/En.MC	65.1	–	–	78.2	–	–	83.4	–	72.1	82.5	–
	NIH/Multi-needle	98.8	–	–	97.5	–	–	98.1	–	100.0	100.0	90.8
Multilingual	MGSM (0-shot, CoT)	68.9	53.2	29.9	86.9	71.1	51.4	91.6	–	85.9	90.5	91.6

According to benchmarks in “Llama 3 herd of models”, Llama 3 8B performs much better than Gemma 2 9B, in 0 shot CoT beating 51.9 and 84.5 in two maths datasets vs 44.3 and 76.7. In this paper, authors state that there was 25% of mathematical and reasoning tokens in the dataset, as compared to Gemma whose dataset prioritises general instruction-following tasks. The dataset includes Code, Wiki, Science and Web (Gemma: Open Models Based on Gemini Research and Technology). However the instruct version we used was trained on mathematics datasets as well (HuggingFace model card screenshot) which helps it perform well on logical reasoning and address mathematical queries. The dataset on which Phi3.5 was trained on had reasoning-dense, heavily filtered datasets, which emphasise knowledge and accuracy, which allows it to perform well on maths-related tasks. On MMLU maths in models we used, LLaMa was 66%, Phi was 68.8%, and Gemma was 64.3%, while on another maths dataset GSM8K LLaMa was 90.3% while Phi was 82.5% and GSM8K was 76.7%



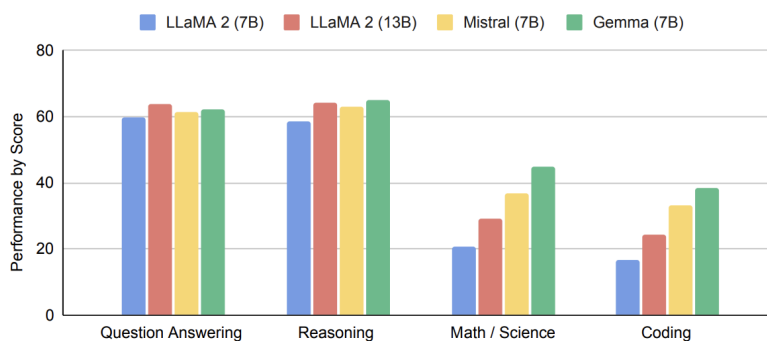
Training Dataset

These models were trained on a dataset of text data that includes a wide variety of sources, totaling 6 trillion tokens. Here are the key components:

- Web Documents: A diverse collection of web text ensures the model is exposed to a broad range of linguistic styles, topics, and vocabulary. Primarily English-language content.
- Code: Exposing the model to code helps it to learn the syntax and patterns of programming languages, which improves its ability to generate code or understand code-related questions.
- Mathematics: Training on mathematical text helps the model learn logical reasoning, symbolic representation, and to address mathematical queries.

Model Size and Complexity:

Llama 3.1 has 8 billion parameters, Phi-3.5 has 3.8 billion parameters, and Gemma has 2 billion. The large number of parameters in Llama allows it to generalise better in zero shot and zero shot chain of thought prompting methods. Due to extensive training, it can handle more complex reasoning and problem-solving tasks, like college mathematics in college mathematics MMLU datasets. The microsoft/Phi-3.5-mini-instruct model that we used at a context length of 128k, which is the same as the Llama model we used. The longer context length helps with better Handling of Complex Inputs, like few-shot or reAct prompting. Its training was focused on improving the size and quality of dataset while keeping the parameters less, hence making it a powerful model which can run on a phone and solve complex mathematical tasks. Gemma-2B uses fewer tokens and has limited optimization for specific domains like maths, and also has very few parameters, limiting its ability to generalise well to complex tasks. Hence resulting in lower performance. According to the paper Gemma: Open Models Based on Gemini Research and Technology, its 7 billion counterpart beats Llama 2 and Mistral in Math and Science tasks. For maths datasets used for benchmarking were GSM8K and MATH. The 2 billion version also beats Llama 2 7 billion in both datasets and 13 billion in MATH dataset. On the MMLU dataset we used, gemma has the highest score of 64.3%, which demonstrates high performance.



Benchmark	metric	LLaMA-2		Mistral		Gemma	
		7B	13B	7B	2B	7B	7B
MMLU	5-shot, top-1	45.3	54.8	62.5	42.3	64.3	
HellaSwag	0-shot	77.2	80.7	81.0	71.4	81.2	
PIQA	0-shot	78.8	80.5	82.2	77.3	81.2	
SIQA	0-shot	48.3	50.3	47.0*	49.7	51.8	
Boolq	0-shot	77.4	81.7	83.2*	69.4	83.2	
Winogrande	partial scoring	69.2	72.8	74.2	65.4	72.3	
CQA	7-shot	57.8	67.3	66.3*	65.3	71.3	
OBQA		58.6	57.0	52.2	47.8	52.8	
ARC-e		75.2	77.3	80.5	73.2	81.5	
ARC-c		45.9	49.4	54.9	42.1	53.2	
TriviaQA	5-shot	72.1	79.6	62.5	53.2	63.4	
NQ	5-shot	25.7	31.2	23.2	12.5	23.0	
HumanEval	pass@1	12.8	18.3	26.2	22.0	32.3	
MBPP [†]	3-shot	20.8	30.6	40.2*	29.2	44.4	
GSM8K	maj@1	14.6	28.7	35.4*	17.7	46.4	
MATH	4-shot	2.5	3.9	12.7	11.8	24.3	
AGIEval		29.3	39.1	41.2*	24.2	41.7	
BBH		32.6	39.4	56.1*	35.2	55.1	
Average		47.0	52.2	54.0	44.9	56.4	

Architecture and Inference Techniques Used:

The llama model uses more advanced inference techniques and optimizations, such as its mixture-of-experts (MoE) architecture, with which it dynamically allocates resources for complex tasks like maths and reasoning. It uses Grouped Query Attention (GQA) technique in its architecture. Phi 3.5 also uses advanced optimizations for efficient reasoning, including longer context windows, but is more limited in its computational power compared to Meta-Llama. It

uses a streamlined transformer decoder with 32 layers and optimised quantization, offering fast inference. Gemma employs Multi-Head Attention and RoPE embeddings, focusing on instruction-following, but it lacks the architectural depth needed for high-level mathematical problem-solving and the deeper optimizations found in Meta-Llama and Phi-3.

Some more benchmarks from <https://huggingface.co/microsoft/Phi-3.5-mini-instruct>:

Benchmark	Phi-3.5 Mini-Ins	Phi-3.0-Mini-128k-Instruct (June2024)	Mistral-7B-Instruct-v0.3	Mistral-Nemo-12B-Ins-2407	Llama-3.1-8B-Ins	Gemma-2-9B-Ins	Gemini 1.5 Flash	GPT-4o-mini-2024-07-18 (Chat)
Multilingual MMLU	55.4	51.08	47.4	58.9	56.2	63.8	77.2	72.9
Multilingual MMLU-Pro	30.9	30.21	15.0	34.0	21.4	43.0	57.9	53.2
MGSM	47.9	41.56	31.8	63.3	56.7	75.1	75.8	81.7
MEGA MLQA	61.7	55.5	43.9	61.2	45.2	54.4	61.6	70.0
MEGA TyDi QA	62.2	55.9	54.0	63.7	54.5	65.6	63.6	81.8
MEGA UDPOS	46.5	48.1	57.2	58.2	54.1	56.6	62.4	66.0
MEGA XCOPA	63.1	62.4	58.8	10.8	21.1	31.2	95.0	90.3
MEGA XStoryCloze	73.5	73.6	75.5	92.3	71.0	87.0	20.7	96.6
Average	55.2	52.3	47.9	55.3	47.5	59.6	64.3	76.6

Long context:

Benchmark	Phi-3.5-mini-instruct	Llama-3.1-8B-instruct	Mistral-7B-instruct-v0.3	Mistral-Nemo-12B-instruct-2407	Gemini-1.5-Flash	GPT-4o-mini-2024-07-18 (Chat)
GovReport	25.9	25.1	26.0	25.6	27.8	24.8
QMSum	21.3	21.6	21.3	22.1	24.0	21.7
Qasper	41.9	37.2	31.4	30.7	43.5	39.8
SQuALITY	25.3	26.2	25.9	25.8	23.5	23.8
SummScreenFD	16.0	17.6	17.5	18.2	16.3	17.0
Average	26.1	25.5	24.4	24.5	27.0	25.4

RULER: a retrieval-based benchmark for long context understanding

Model	4K	8K	16K	32K	64K	128K	Average
Phi-3.5-mini-instruct	94.3	91.1	90.7	87.1	78.0	63.6	84.1
Llama-3.1-8B-instruct	95.5	93.8	91.6	87.4	84.7	77.0	88.3
Mistral-Nemo-12B-instruct-2407	87.8	87.2	87.7	69.0	46.8	19.0	66.2

RepoQA: a benchmark for long context code understanding

Model	Python	C++	Rust	Java	TypeScript	Average
Phi-3.5-mini-instruct	86	67	73	77	82	77
Llama-3.1-8B-instruct	80	65	73	76	63	71
Mistral-7B-instruct-v0.3	61	57	51	61	80	62

<https://ar5iv.labs.arxiv.org/html/2404.14219>

<https://training.continuumlabs.ai/models/foundation-models/analysis-of-llama-3>

<https://arxiv.org/pdf/2407.21783>

<https://arxiv.org/pdf/2403.08295>

MATH dataset: <https://arxiv.org/pdf/2103.03874>

<https://anakin.ai/blog/llama-3-vs-phi-3/>

