Llama3 & DeepSeek v3 (part 2)

Architecture | Inference

Dr. Karthik Mohan, March 3rd 2025

Today's Talk

1. Llama3 Arch

2. DeepSeekV3 Arch

3. Benchmarking out of box ✓

MP3

X 4. Notebook Walkthrough

MP3

Part 1 :-

- Llama3 Fine-tuning on dataset
- Kaggle Contest

Task :- Intent Detection

- Simple
  - Single Intent
- Complex
  - More than one intent

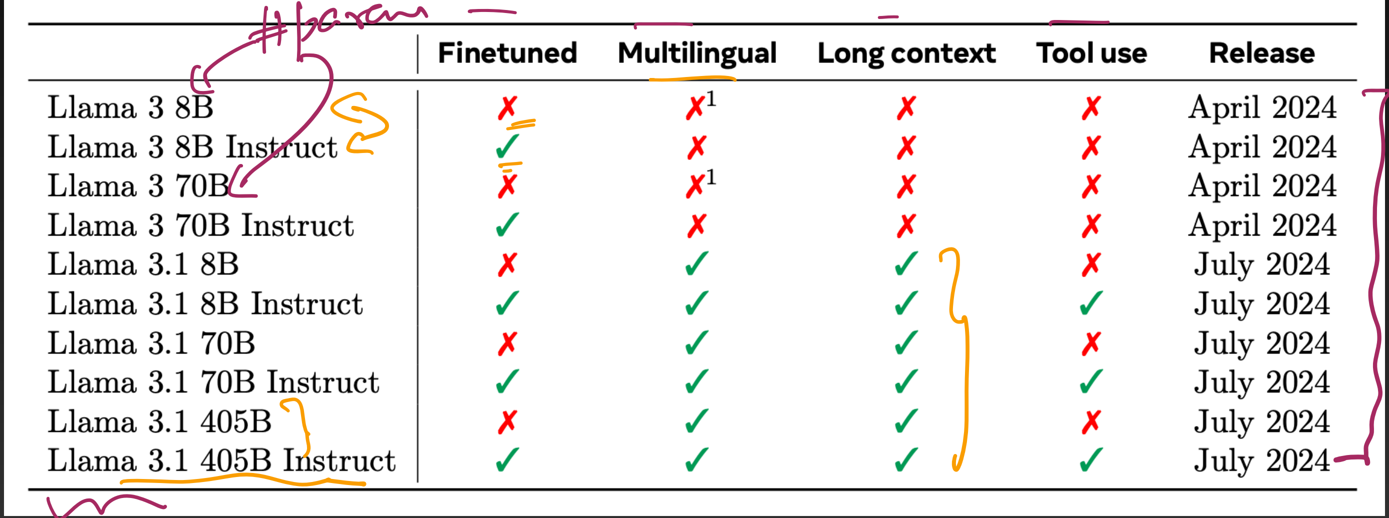
Part 2 : Stable Diffusion based assignment (1week)

Llama3 Herd

Herd of models including 405B LM, 70B, 8B, 1B versions and also Llama Guard 3 for input/output safety

Reference: https://arxiv.org/pdf/2407.21783

Llama 3 Herd of Models



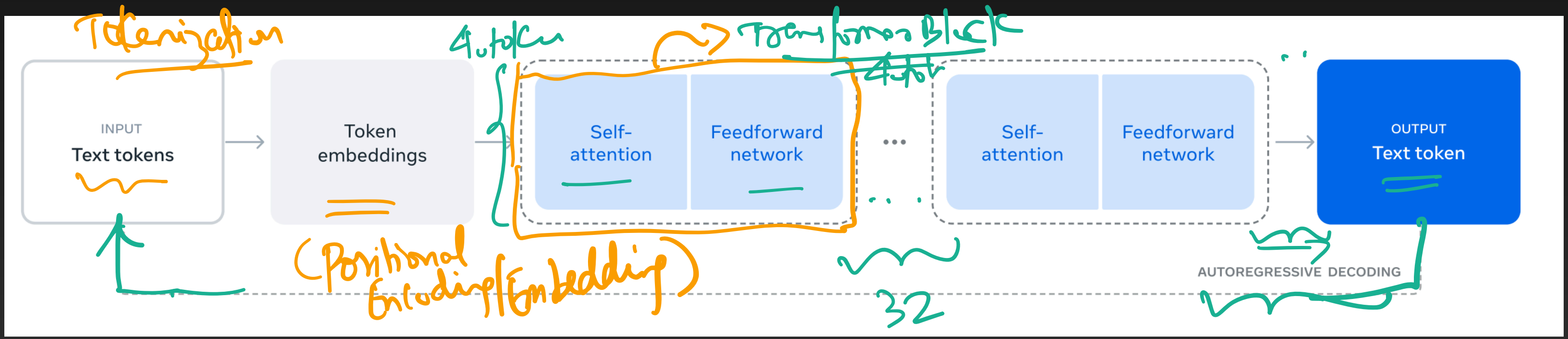
	Finetuned	Multilingual	Long context	Tool use	Release
Llama 3 8B	✗	✗[1]	✗	✗	April 2024
Llama 3 8B Instruct	✓	✗	✗	✗	April 2024
Llama 3 70B	✗	✗[1]	✗	✗	April 2024
Llama 3 70B Instruct	✓	✗	✗	✗	April 2024
Llama 3.1 8B	✗	✓	✓	✗	July 2024
Llama 3.1 8B Instruct	✓	✓	✓	✓	July 2024
Llama 3.1 70B	✗	✓	✓	✗	July 2024
Llama 3.1 70B Instruct	✓	✓	✓	✓	July 2024
Llama 3.1 405B	✗	✓	✓	✗	July 2024
Llama 3.1 405B Instruct	✓	✓	✓	✓	July 2024

#param

Reference: https://arxiv.org/pdf/2407.21783

92 page Paper

Llama3 Architecture



INPUT
Text tokens

Token embeddings

Self-attention

Feedforward network

...

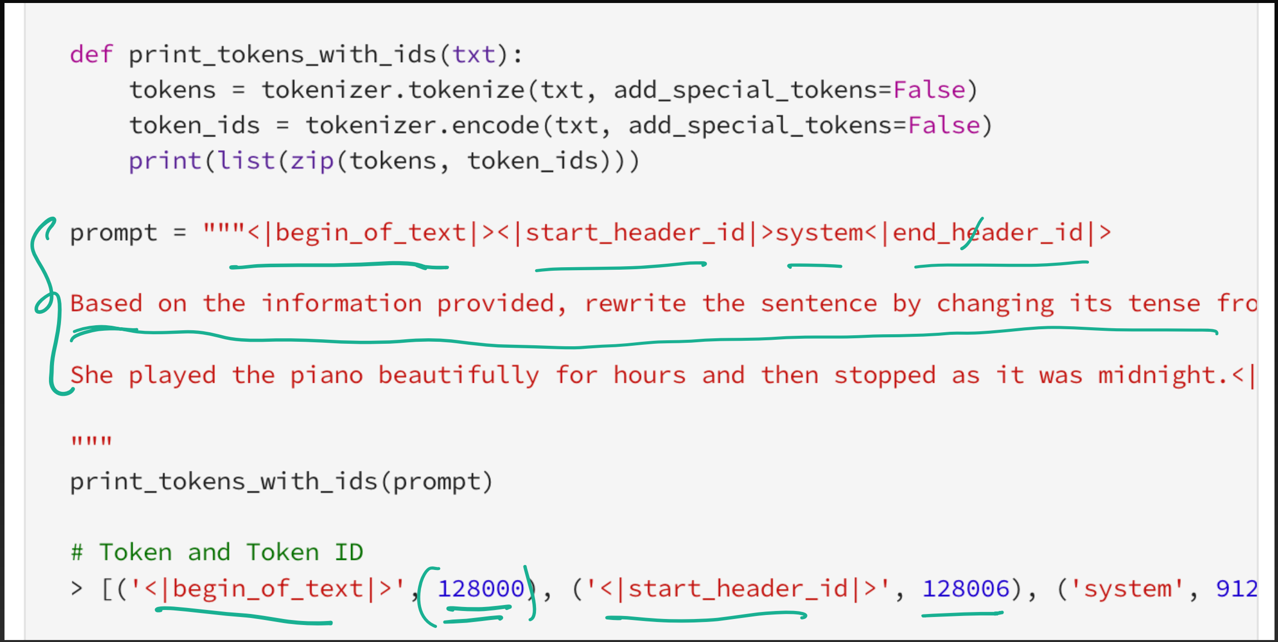
Self-attention

Feedforward network

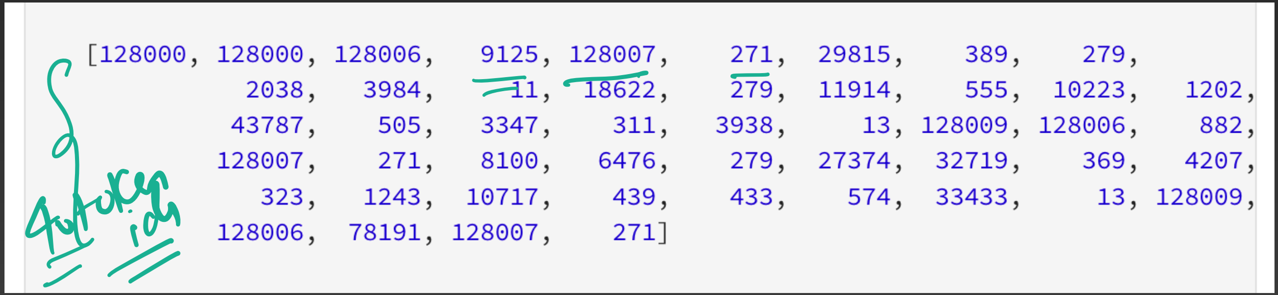
OUTPUT
Text token

AUTOREGRESSIVE DECODING

Tokenization



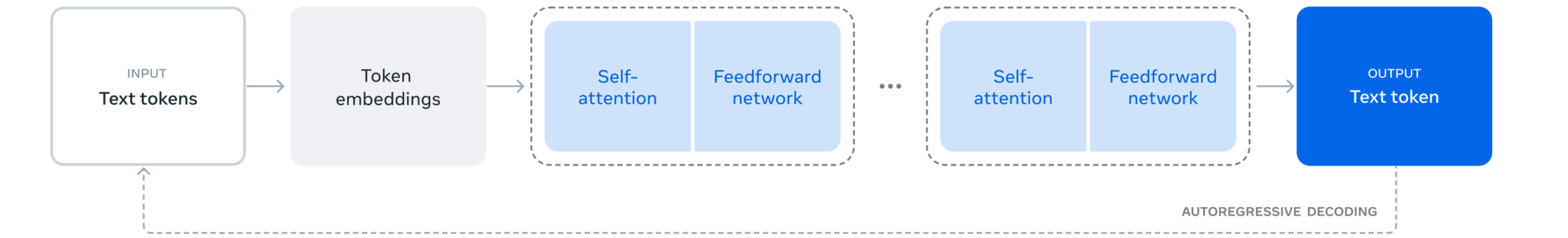
```
def print_tokens_with_ids(txt):
    tokens = tokenizer.tokenize(txt, add_special_tokens=False)
    token_ids = tokenizer.encode(txt, add_special_tokens=False)
    print(list(zip(tokens, token_ids)))

prompt = """<|begin_of_text|><|start_header_id|>system<|end_header_id|>

Based on the information provided, rewrite the sentence by changing its tense fro

She played the piano beautifully for hours and then stopped as it was midnight.<|

"""
print_tokens_with_ids(prompt)

# Token and Token ID
> [('<|begin_of_text|>', 128000), ('<|start_header_id|>', 128006), ('system', 912
```



```
[128000, 128000, 128006,   9125, 128007,    271,  29815,    389,    279,
           2038,   3984,     11,  18622,    279,  11914,    555,  10223,   1202,
          43787,    505,   3347,    311,   3938,     13, 128009, 128006,    882,
         128007,    271,   8100,   6476,    279,  27374,  32719,    369,   4207,
            323,   1243,  10717,    439,    433,    574,  33433,     13, 128009,
         128006,  78191, 128007,    271]
```

49 token ids

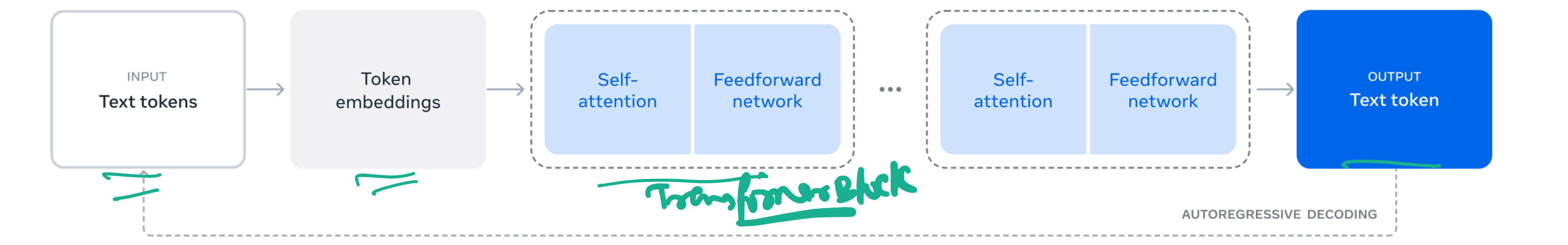
ICE #1



Assume that Llama4 is trained on 40T tokens of data. It has a context window of 256k tokens and has a tokenizer vocab size of 108k tokens and each token has a token embedding size of 4096.
What is the number of classes present in the classifier of the LM head to generate the next token in auto-regressive decoding?

a) 256k
b) 40T
c) [✓ 108k]
d) 128k
e) 4096

Llama3 Architecture



INPUT
Text tokens

Token embeddings

Self-attention

Feedforward network

...

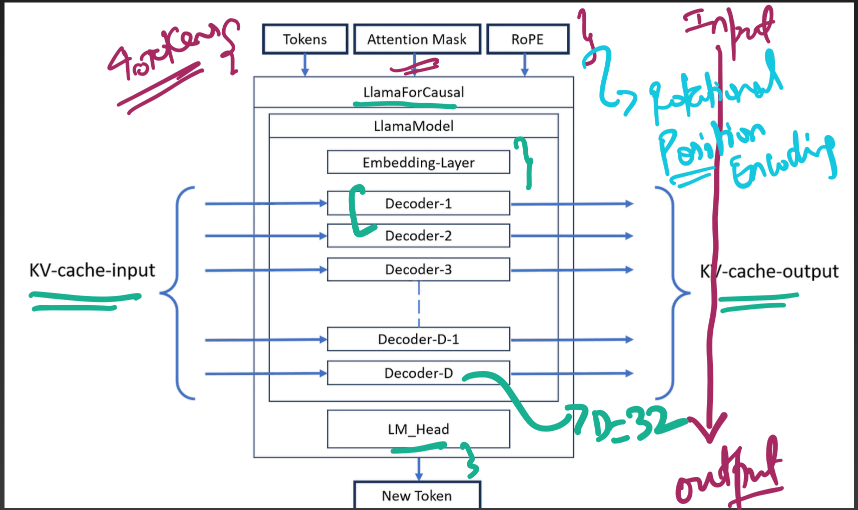
Self-attention

Feedforward network

OUTPUT
Text token

AUTOREGRESSIVE DECODING

Transformer Block



4 tokens

Tokens

Attention Mask

RoPE

LlamaForCausal

LlamaModel

Embedding-Layer

Decoder-1

Decoder-2

Decoder-3

Decoder-D-1

Decoder-D

LM_Head

New Token

KV-cache-input

KV-cache-output

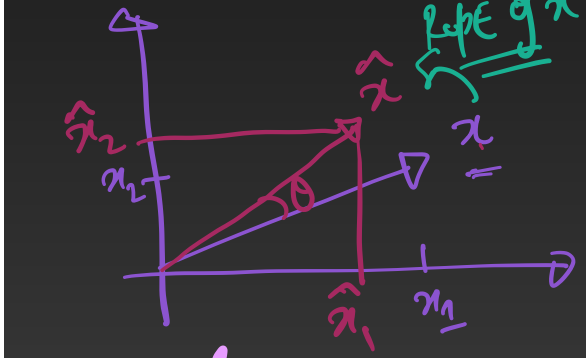
Input

Rotational Position Encoding

D=32

output

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \to \text{Rotation matrix}$$



RoPE of x

Previously

$Emb(token) = T(n) + Pos(n)$

Now

$R_\theta(T(n))$

$$R_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \qquad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 2d$$

token embedding

$$\|x\|_2 = \sqrt{x_1^2+x_2^2}$$

$$\hat{x} = R_\theta x = \begin{bmatrix} x_1\cos\theta - x_2\sin\theta \\ x_1\sin\theta + x_2\cos\theta \end{bmatrix}$$

$$\|\hat{x}\|_2^2 = (x_1\cos\theta - x_2\sin\theta)^2 + (x_1\sin\theta + x_2\cos\theta)^2$$

$$= \underline{x_1^2\cos^2\theta} + \underline{x_2^2\sin^2\theta} - \underline{2x_1x_2\cos\theta\sin\theta} + \underline{x_1^2\sin^2\theta} + \underline{x_2^2\cos^2\theta} + \underline{2x_1x_2\cos\theta\sin\theta}$$

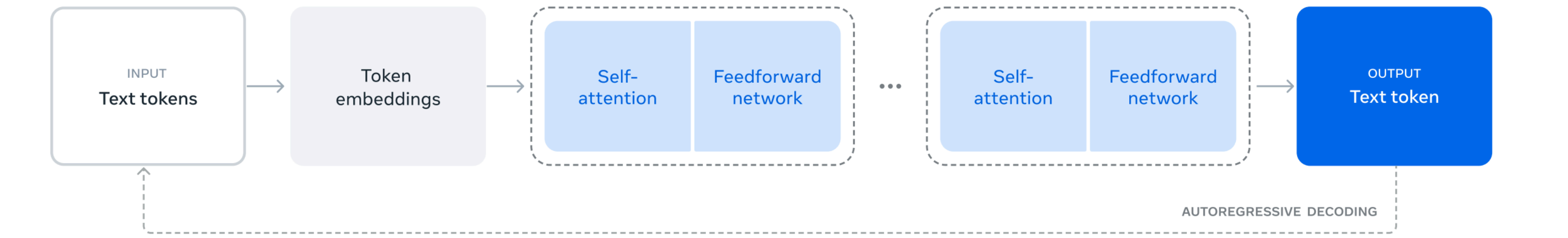
$$= x_1^2\underline{(\cos^2\theta+\sin^2\theta)} + x_2^2\underline{(\sin^2\theta+\cos^2\theta)} + 0$$

$$= x_1^2+x_2^2$$

$$\|\hat{x}\|_2 = \sqrt{x_1^2+x_2^2} = \|x\|_2$$

Less rotation in RoPE

ICE #2



Assume that input sentence has 100 words and tokenized into 150 tokens. The 150 tokens are now assigned a token embedding and passed through 64 decoder blocks of Llama model. At the very end, a new token is also generated. How many tokens exist right after the 64 decoder blocks and how many new tokens are generated?

a) 150,64

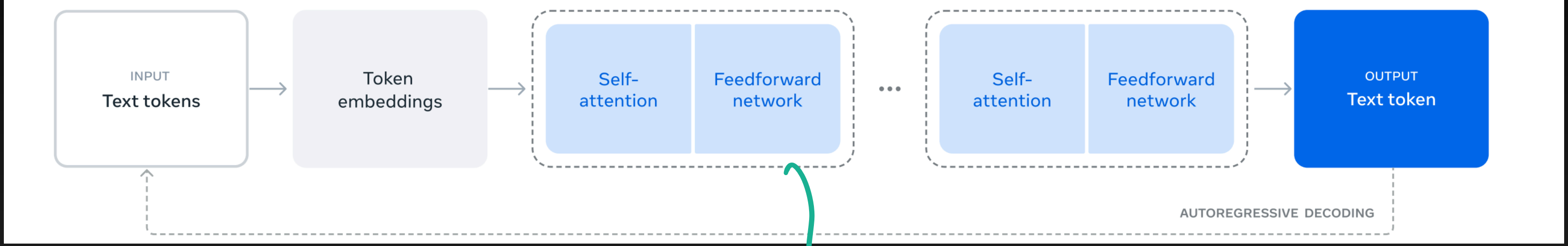
b) 64,1

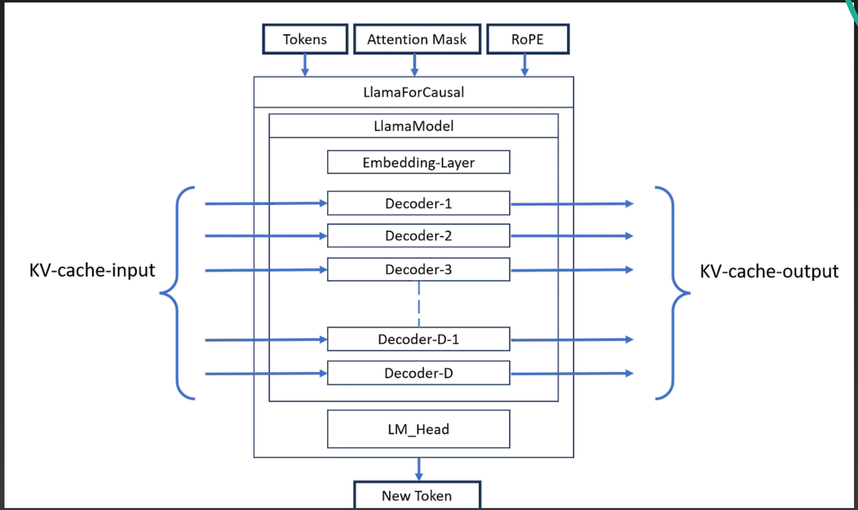
c) 1,1

d) 150,1 ✓

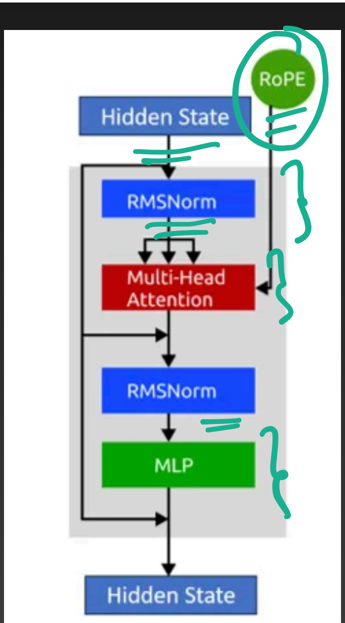
e) 150,150

Llama3 Architecture

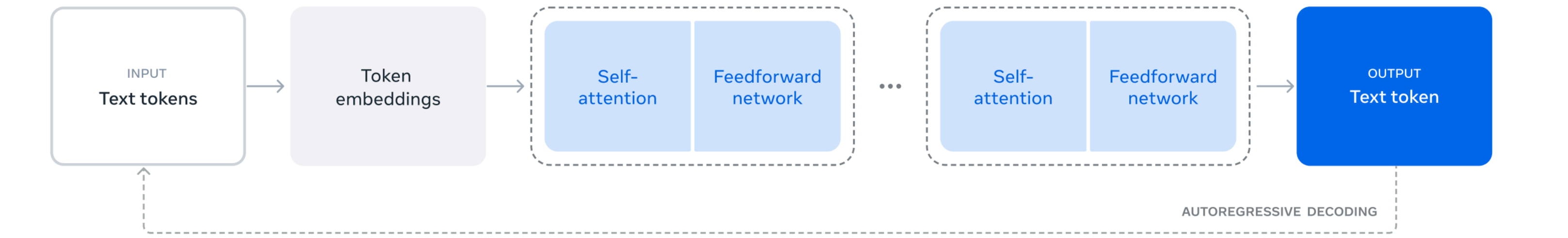




Transformer Block (Decoder)



ICE #3



Assume that each word in a sentence corresponds to a token (for simplicity). Consider the sentence: “The sun rises”. Passing this into llama produces the output token as “in”. How many tokens will be passed in to the next step of autoregressive decoding and what will the expected output?

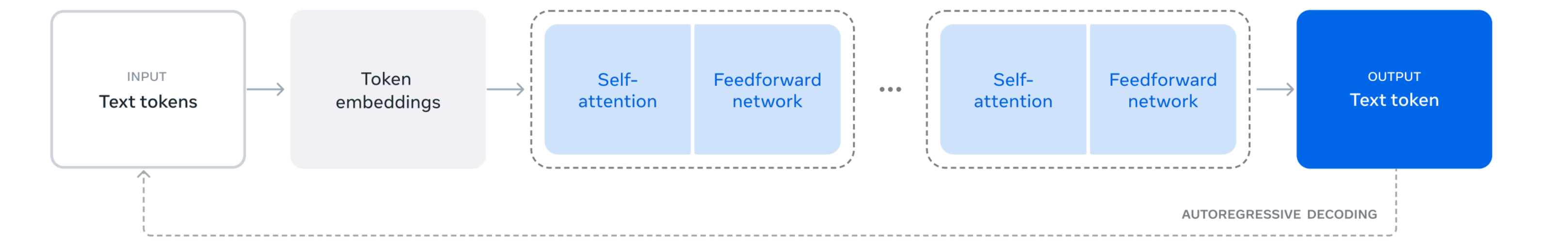
a) 3,”east”

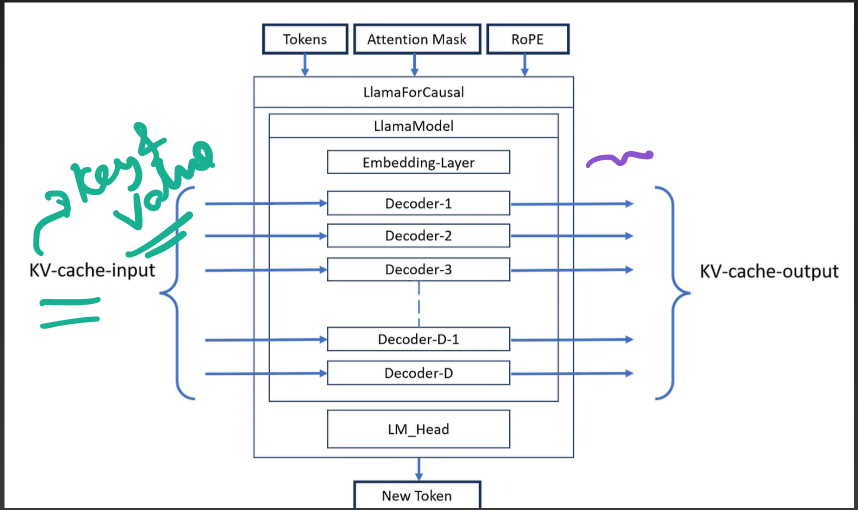
b) 4,”east”

c) 3,”the

[✓ d) 4, “the”]

Llama3 Architecture





Gen = 0

40 token Input

1 new token output

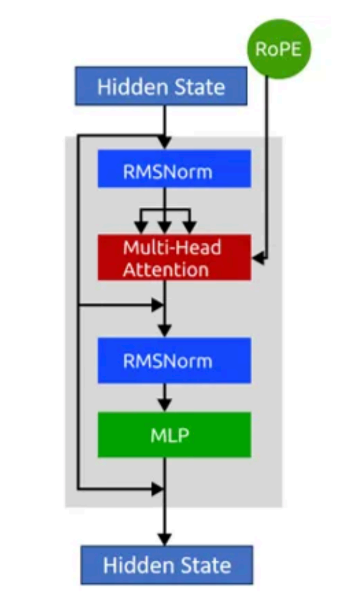
Gen=1

41 token Input

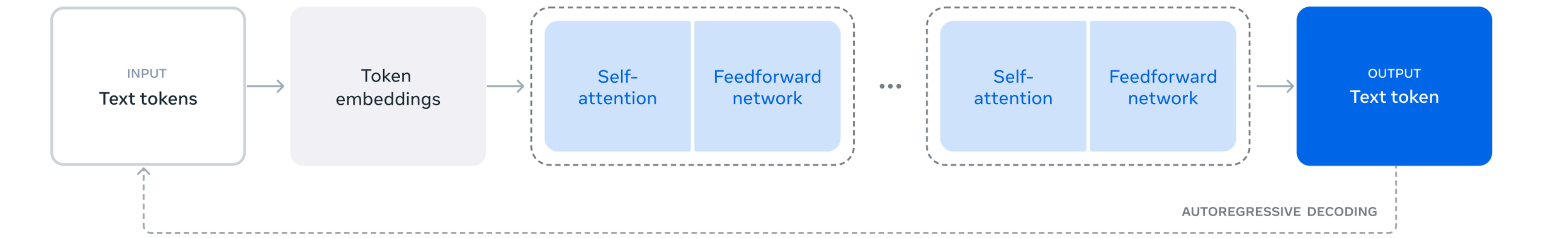
1 new output

Gen=2

..



ICE #4



What is the role of KV-cache in the Llama architecture?

a) To compute attention using query, keys and values

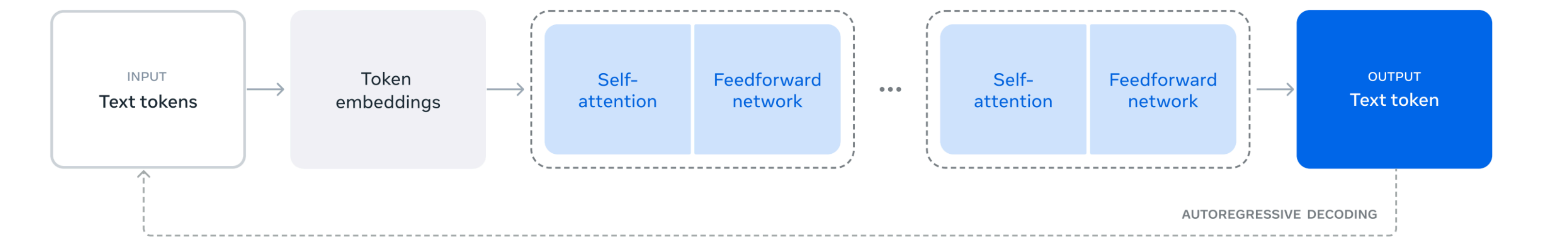
b) To cache all the model parameters in memory ✗

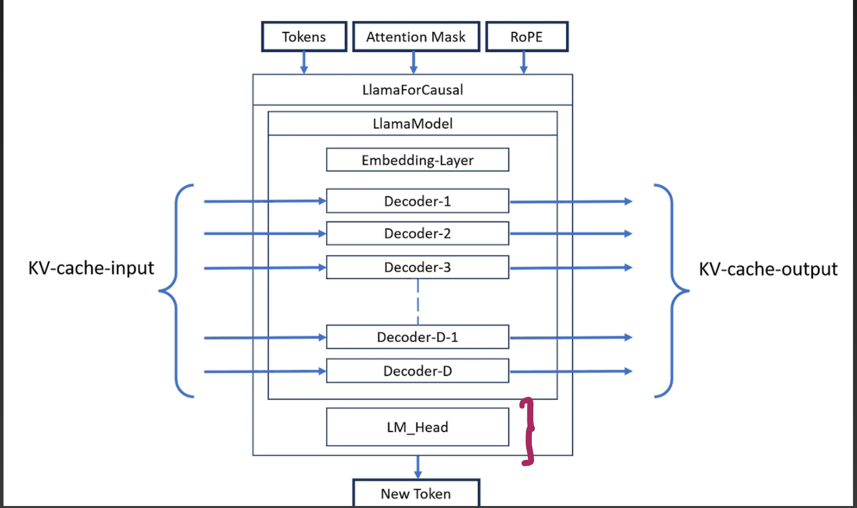
c) To speed up computation in the auto-regressive decoding process ✓

d) To cache intermediate query, token embeddings to be used in the next decoding phase ✓

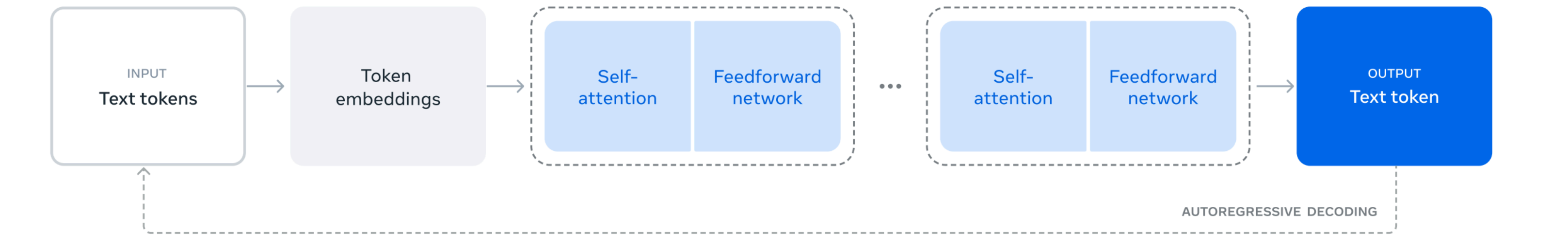
e) All of the above ✗

Llama3 Architecture





ICE #5 | LM head



What is the composition of LM head?

a) Linear transformation + softmax activation ← Probabilities

b) Non-linear transformation + relu activation

c) Non-linear transformation + softmax activation

d) Linear transformation + relu activation

Llama3 Key Features

Context Window: 128k tokens (4k tokens for Llama2)

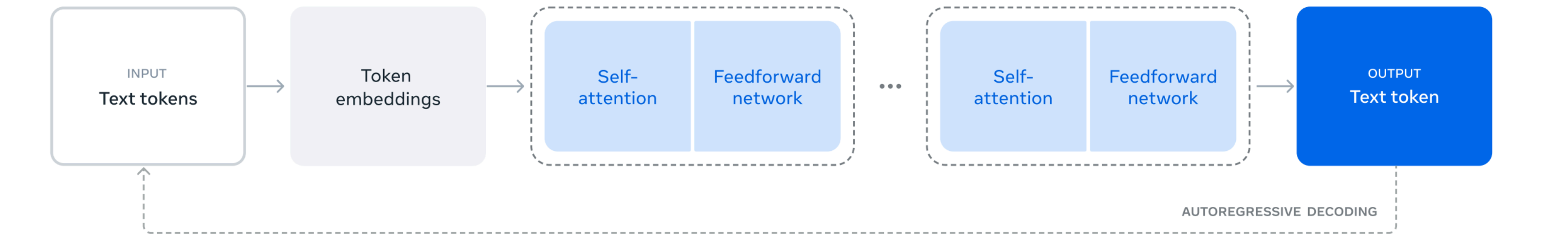
Vocab size: 128k tokens

Training data: 15T tokens

Decoder blocks: 32 (Transformer blocks)

Positional Embedding: RoPE

ICE #6


INPUT
Text tokens

Token embeddings

Self-attention

Feedforward network

...

Self-attention

Feedforward network

OUTPUT
Text token

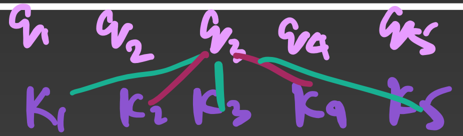
AUTOREGRESSIVE DECODING

Why is increasing the input context window been a challenge for LLM models. Only recently have LLM models increased it from 4000 tokens to 100k tokens

a) Compute increases linearly with context window size

b) Compute increases quadratically with context window size ✓

c) Compute increases exponentially with context window size



O(25) compute

T token → $O(T^2)$

Llama2 vs Llama3

7 times larger pre-train data set. **15 Trillion Tokens** of data ~ 150 million books
High-quality filters to filter out bad data in training - Use Llama2 } Reverse Distillation
Better "data mix" - Trivia, STEM, coding, historical knowledge

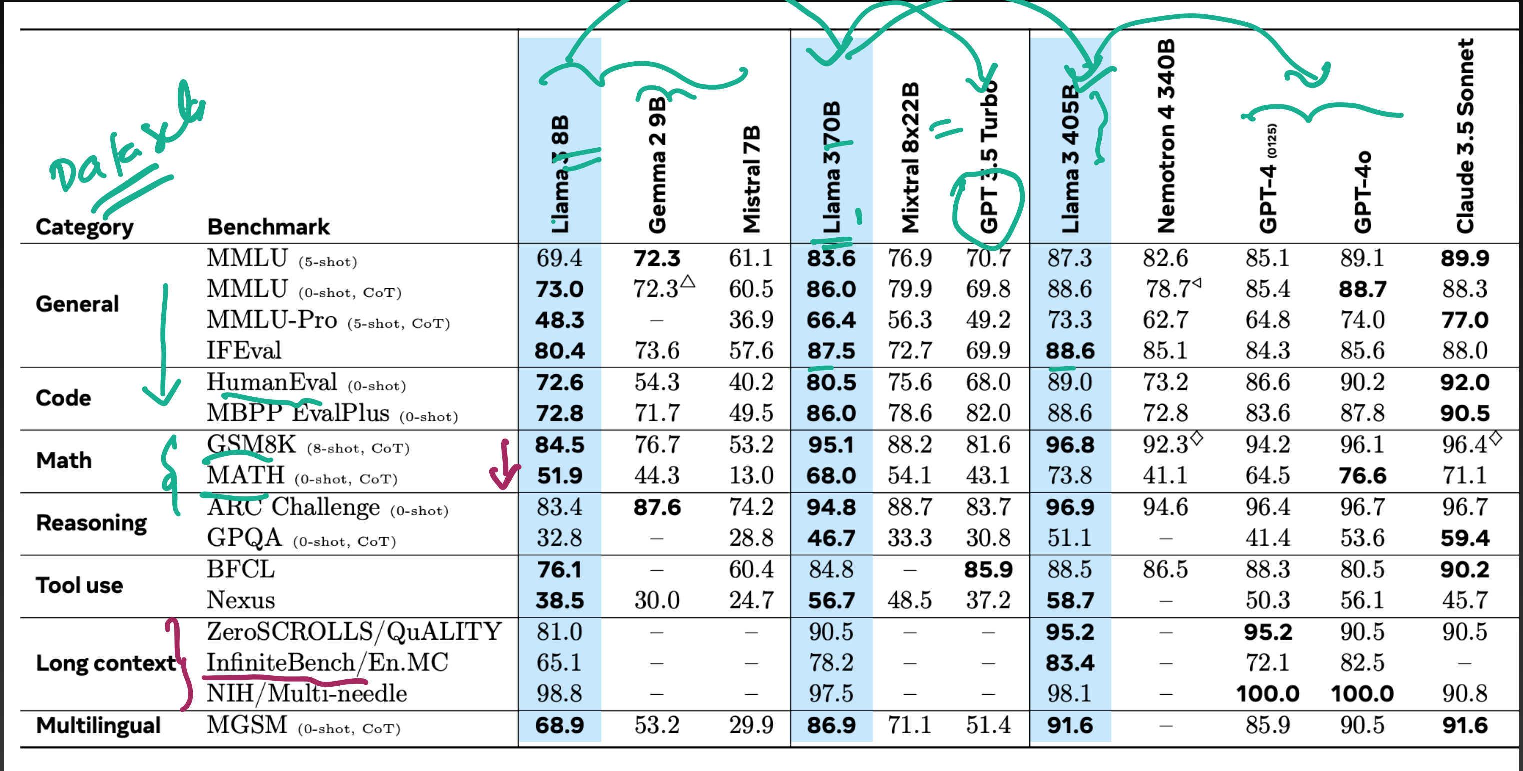
Larger model means better performance (8B vs 70B)
But more data = better performance (also avoids over-fitting). Log-linear improvement from 200B to 15T tokens

Llama3 vs DeepSeek

	Llama3	DeepSeek V3
Parameters	405b	405b with 37b active at inference
Architecture	(2017) Traditional Transformers (Decoder)	Transformer with MOE and MLA
Context Length	128k tokens	128k tokens
Post Training	Instruct FT	Instruct FT
RL	DPO	DPO

Mix of Experts make this possible

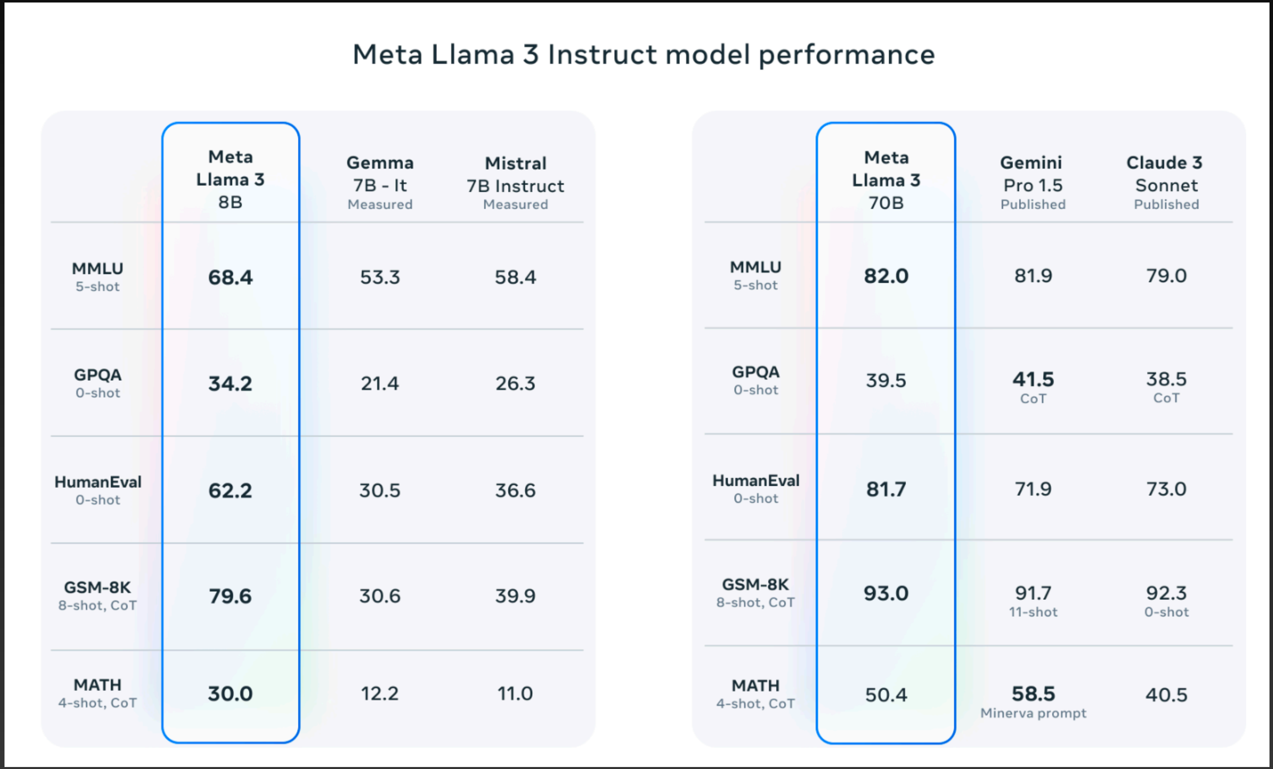
Llama3 Benchmarks



Dataset

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-4 (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**

Llama 3 Instruct Performance



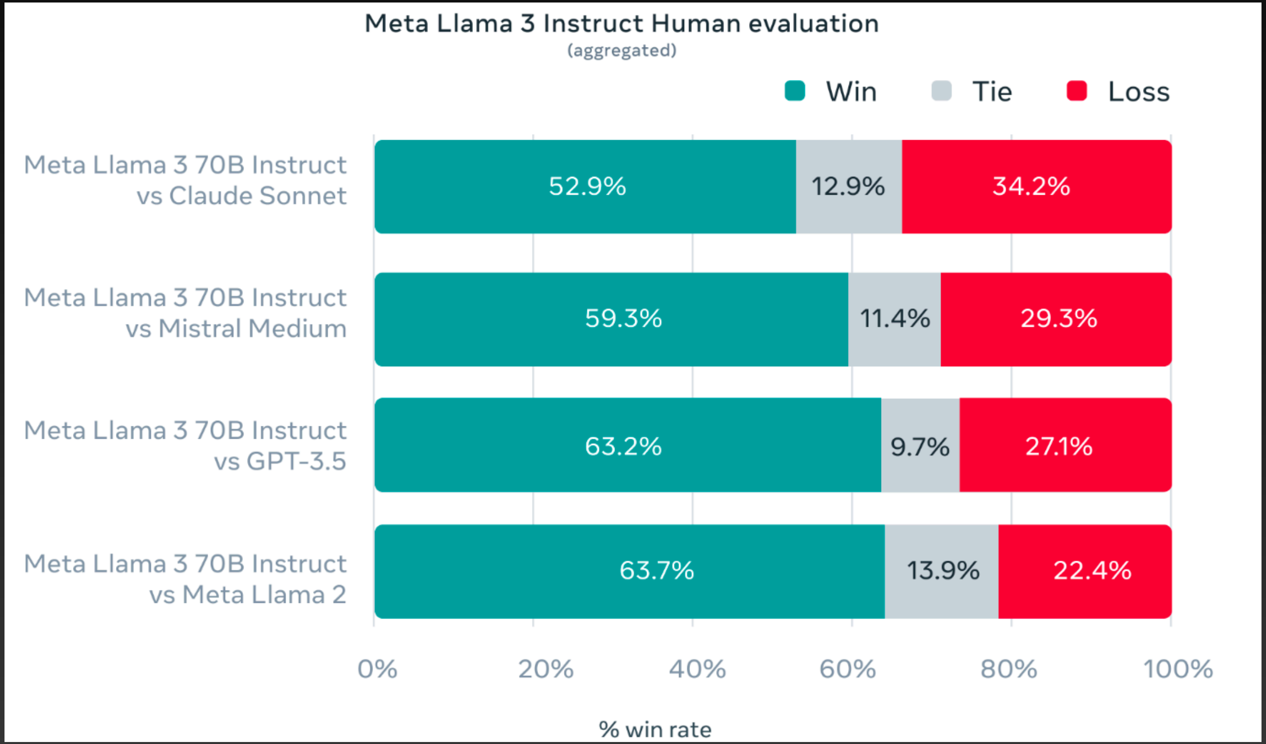
Meta Llama 3 Instruct model performance

	Meta Llama 3 8B	Gemma 7B - It Measured	Mistral 7B Instruct Measured
MMLU 5-shot	**68.4**	53.3	58.4
GPQA 0-shot	**34.2**	21.4	26.3
HumanEval 0-shot	**62.2**	30.5	36.6
GSM-8K 8-shot, CoT	**79.6**	30.6	39.9
MATH 4-shot, CoT	**30.0**	12.2	11.0

	Meta Llama 3 70B	Gemini Pro 1.5 Published	Claude 3 Sonnet Published
MMLU 5-shot	**82.0**	81.9	79.0
GPQA 0-shot	39.5	**41.5** CoT	38.5 CoT
HumanEval 0-shot	**81.7**	71.9	73.0
GSM-8K 8-shot, CoT	**93.0**	91.7 11-shot	92.3 0-shot
MATH 4-shot, CoT	50.4	**58.5** Minerva prompt	40.5

Reference: https://ai.meta.com/blog/meta-llama-3/

Llama 3 Instruct Performance



Reference: https://ai.meta.com/blog/meta-llama-3/

DeepSeek V3