

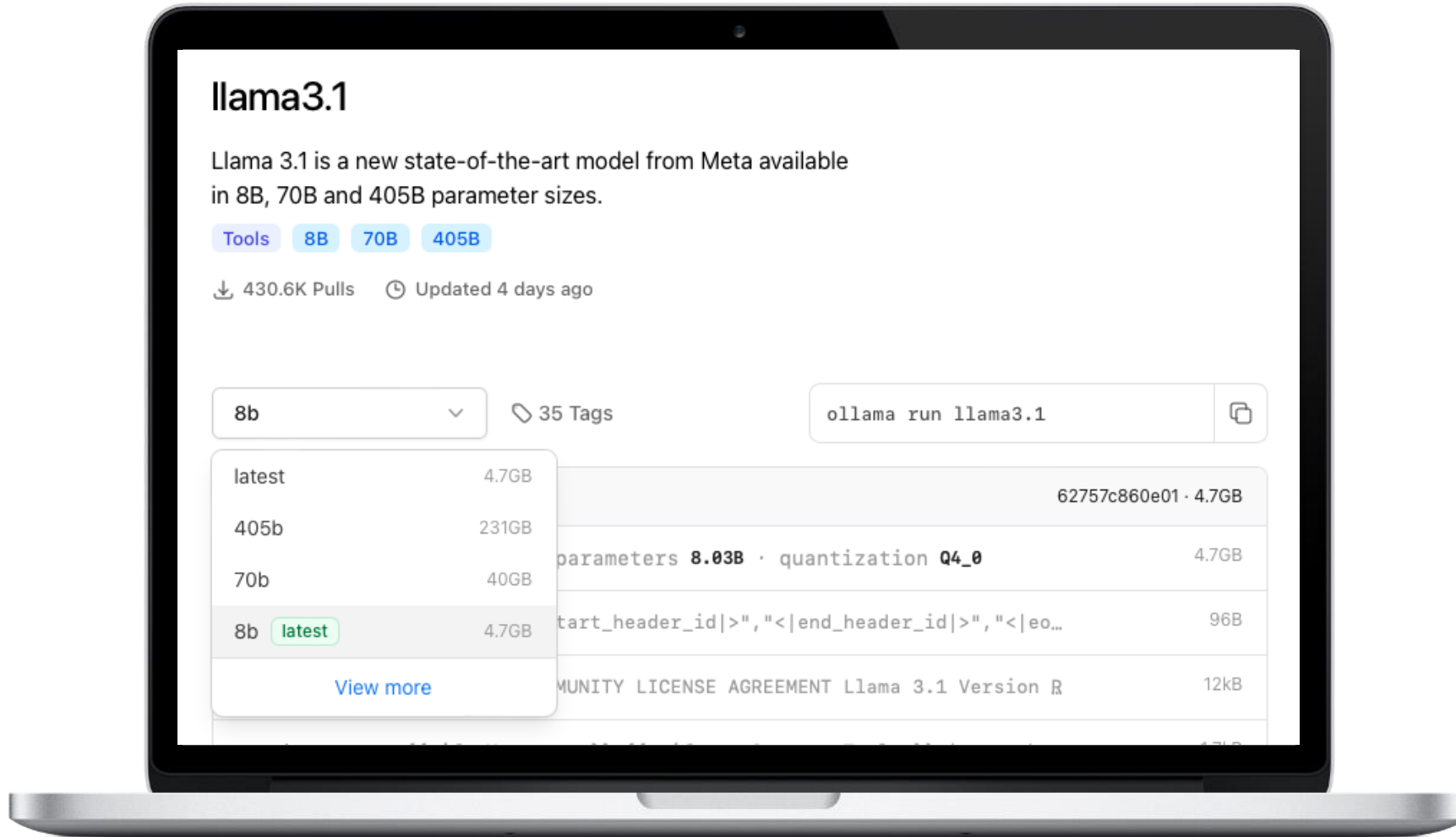
Create Your Own Generative AI Locally

Using LLAMA Developed by Meta

Dr. Eng. Pujiyanto Yugopuspito, M.Sc.



Llama 3.1

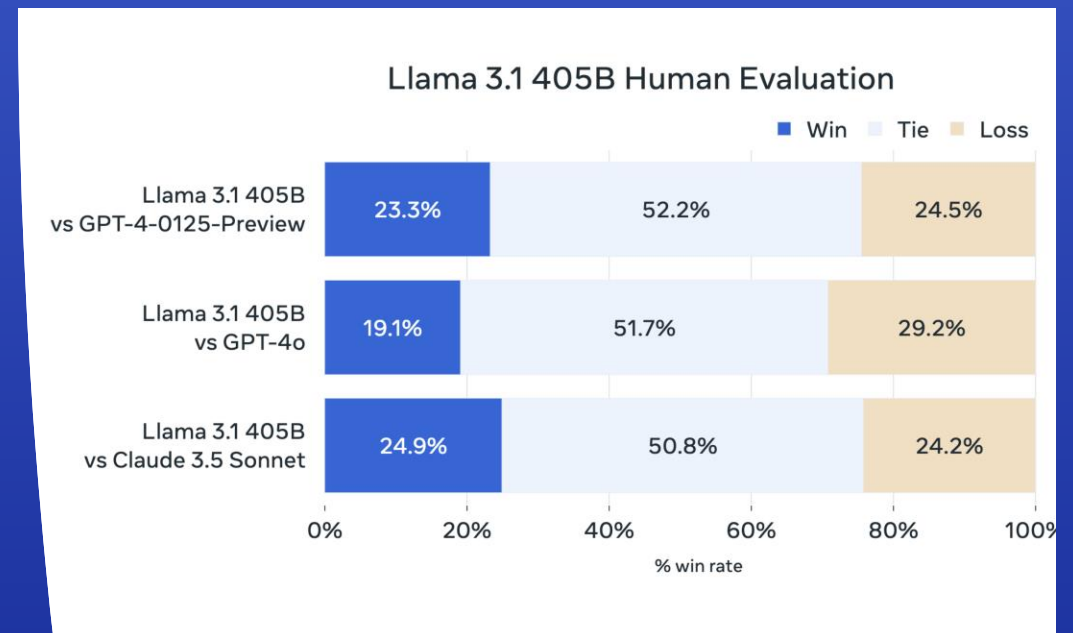


<https://ollama.com>

Llama 3.1 Model Evaluation

- Meta has evaluated the performance on over 150 benchmark datasets that span a wide range of languages, with competing models in real-world scenarios.
- Meta's experimental evaluation suggests that Llama 3.1 flagship model is competitive with leading foundation models across a range of tasks, including GPT-4, GPT-4o, and Claude 3.5 Sonnet.

Category/Benchmark	Llama 3.1 405B	Nemotron 4 340B Instruct	GPT-4 (0125)	GPT-4 Omni	Claude 3.5 Sonnet
General					
MMLU (0-shot, CoT)	88.6	78.7 (non-CoT)	85.4	88.7	88.3
MMLU PRO (5-shot, CoT)	73.3	62.7	64.8	74.0	77.0
IFEval	88.6	85.1	84.3	85.6	88.0
Code					
HumanEval (0-shot)	89.0	73.2	86.6	90.2	92.0
MBPP EvalPlus (base) (0-shot)	88.6	72.8	83.6	87.8	90.5
Math					
GSM8K (0-shot, CoT)	96.8	92.3 (0-shot)	94.2	96.1	96.4 (0-shot)
MATH (0-shot, CoT)	73.8	41.1	64.5	76.6	71.1
Reasoning					
ARC Challenge (0-shot)	96.9	94.6	96.4	96.7	96.7
GPQA (0-shot, CoT)	51.1	-	41.4	53.6	59.4
Tool use					
BFCL	88.5	86.5	88.3	80.5	90.2
Nexus	58.7	-	50.3	56.1	45.7
Long context					
ZeroSCROLLS/QuALITY	95.2	-	95.2	90.5	90.5
InfiniteBench/En,MC	83.4	-	72.1	82.5	-



Llama 3.1 Human Evaluation for Smaller Model

• • •

The higher its value, the better

Category Benchmark	Llama 3.1 8B	Gemma 2 9B IT	Mistral 7B Instruct	Llama 3.1 70B	Mixtral 8x22B Instruct	GPT 3.5 Turbo
General						
MMLU (0-shot, CoT)	73.0	72.3 <small>(5-shot, non-CoT)</small>	60.5	86.0	79.9	69.8
MMLU PRO (5-shot, CoT)	48.3	-	36.9	66.4	56.3	49.2
IFEval	80.4	73.6	57.6	87.5	72.7	69.9
Code						
HumanEval (0-shot)	72.6	54.3	40.2	80.5	75.6	68.0
MBPP EvalPlus <small>(base) (0-shot)</small>	72.8	71.7	49.5	86.0	78.6	82.0
Math						
GSM8K (8-shot, CoT)	84.5	76.7	53.2	95.1	88.2	81.6
MATH (0-shot, CoT)	51.9	44.3	13.0	68.0	54.1	43.1
Reasoning						
ARC Challenge (0-shot)	83.4	87.6	74.2	94.8	88.7	83.7
GPQA (0-shot, CoT)	32.8	-	28.8	46.7	33.3	30.8
Tool use						
BFCL	76.1	-	60.4	84.8	-	85.9
Nexus	38.5	30.0	24.7	56.7	48.5	37.2
Long context						
ZeroSCROLLS/QuALITY	81.0	-	-	90.5	-	-
InfiniteBench/En.MC	65.1	-	-	78.2	-	-
NIH/Multi-needle	98.8	-	-	97.5	-	-
Multilingual						
Multilingual MGSM <small>(0-shot)</small>	68.9	53.2	29.9	86.9	71.1	51.4



Llama 3.1 First Trial

```
>>> I am a newbie in Generative AI, can you describe what is llama?  
You're new to the world of Generative AI!
```

```
Llama stands for "Large Language Model Application" and it's actually a model developed by Meta (the parent company of Facebook).  
It's a type of AI designed to understand and respond to human language in a conversational way.
```

```
Llama is trained on a massive dataset of text, which allows it to generate coherent and context-specific responses to questions  
or prompts. In other words, it can have a conversation with you!
```

```
Here are some key features that make Llama interesting:
```

1. **Conversational**: Llama is designed to understand the nuances of human language, including idioms, colloquialisms, and even humor.
2. **Contextual**: It can keep track of the context within a conversation, allowing it to provide more accurate and relevant responses.
3. **Open-ended**: Llama can respond to open-ended questions or prompts that don't have a clear "right" answer.

```
Llama is not the only AI model out there (e.g., BERT, RoBERTa), but it's one of the most well-known and widely used in  
applications like chatbots, customer support systems, and even language translation tools!
```

```
Do you want to know more about Llama or Generative AI in general?
```

Video Action

Llama is trained on a massive dataset of text, which allows it to generate coherent and context-specific responses to questions or prompts. In other words, it can have a conversation with you!

Here are some key features that make Llama interesting:

1. **Conversational**: Llama is designed to understand the nuances of human language, including idioms, colloquialisms, and even humor.
2. **Contextual**: It can keep track of the context within a conversation, allowing it to provide more accurate and relevant responses.
3. **Open-ended**: Llama can respond to open-ended questions or prompts that don't have a clear "right" answer.

Llama is not the only AI model out there (e.g., BERT, RoBERTa), but it's one of the most well-known and widely used in applications like chatbots, customer support systems, and even language translation tools!

Do you want to know more about Llama or Generative AI in general?

>>>

The background is a solid blue color. In the top right and bottom left corners, there are large, thick, orange curved shapes that resemble parts of a circle or a stylized 'C' shape.

THANK YOU

Universitas Pelita Harapan