

Phi-3

Phi-3

- 1. Model size
- 2. Performance
- 3. Pull Phi-3
- 4. Use Phi-3
 - 4.1. Run Phi-3
 - 4.2. Have a conversation
 - 4.3. End the conversation
- References

Demo Environment

Development board: Jetson Orin series motherboard

SSD: 128G

Tutorial application scope: Whether the motherboard can run is related to the available memory of the system. The user's own environment and the programs running in the background may cause the model to fail to run.

Motherboard model	Run directly with Ollama	Run with Open WebUI
Jetson Orin NX 16GB	√	√
Jetson Orin NX 8GB	√	√
Jetson Orin Nano 8GB	√	√
Jetson Orin Nano 4GB	√	√

Phi-3 is a powerful, cost-effective small language model (SLM) from Microsoft that outperforms models of the same and higher sizes in a variety of language, reasoning, encoding, and math benchmarks.

1. Model size

Model	Parameters
Phi-3 (Mini)	3.8B
Phi-3 (Medium)	14B

2. Performance

Category	Benchmark	Phi-3				Gemma-7b	Mistral-7b	Mixtral-8x7b	Llama-3-8B-In	GPT3.5-Turbo-1106	Claude-3 Sonnet
		Phi-3-Mini-4K-In	Phi-3-Mini-128K-In	Phi-3-Small (Preview)	Phi-3-Medium (Preview)						
Popular Aggregate Benchmarks	AGI Eval (0-shot)	37.5	36.9	45	48.4	42.1	35.1	45.2	42	48.4	48.4
	MMLU (5-shot)	68.8	68.1	75.6	78.2	63.6	61.7	70.5	66.5	71.4	73.9
	BigBench Hard (0-shot)	71.7	71.5	74.9	81.3	59.6	57.3	69.7	51.5	68.3	--
Language Understanding	ANLI (7-shot)	52.8	52.8	55	58.7	48.7	47.1	55.2	57.3	58.1	68.6
	Hellaswag (5-shot)	76.7	74.5	78.7	83	49.8	58.5	70.4	71.1	78.8	79.2
Reasoning	ARC Challenge (10-shot)	84.9	84	90.7	91	78.3	78.6	87.3	82.8	87.4	91.6
	ARC Easy (10-shot)	94.6	95.2	97.1	97.8	91.4	90.6	95.6	93.4	96.3	97.7
	BoolQ (0-shot)	77.6	78.7	82.9	86.6	66	72.2	76.6	80.9	79.1	87.1
	CommonsenseQA (10-shot)	80.2	78	80.3	82.6	76.2	72.6	78.1	79	79.6	82.6
	MedQA (2-shot)	53.8	55.3	58.2	69.4	49.6	50	62.2	60.5	63.4	67.9
	OpenBookQA (10-shot)	83.2	80.6	88.4	87.2	78.6	79.8	85.8	82.6	86	90.8
	PIQA (5-shot)	84.2	83.6	87.8	87.7	78.1	77.7	86	75.7	86.6	87.8
	Social IQA (5-shot)	76.6	76.1	79	80.2	65.5	74.6	75.9	73.9	68.3	80.2
	TruthfulQA (MC2) (10-shot)	65	63.2	68.7	75.7	52.1	53	60.1	63.2	67.7	77.8
	WinoGrande (5-shot)	70.8	72.5	82.5	81.4	55.6	54.2	62	65	68.8	81.4
Factual Knowledge	TriviaQA (5-shot)	64	57.1	59.1	75.6	72.3	75.2	82.2	67.7	85.8	65.7
Math	GSM8K Chain of Thought (0-shot)	82.5	83.6	88.9	90.3	59.8	46.4	64.7	77.4	78.1	79.1
Code generation	HumanEval (0-shot)	59.1	57.9	59.1	55.5	34.1	28	37.8	60.4	62.2	65.9
	MBPP (3-shot)	53.8	62.5	71.4	74.5	51.5	50.8	60.2	67.7	77.8	79.4

3. Pull Phi-3

Using the pull command will automatically pull the model from the Ollama model library:

```
ollama pull phi3:3.8b
```

```

jetson@yahboom: ~
jetson@yahboom:~$ ollama pull phi3:3.8b
pulling manifest
pulling 633fc5be925f... 100% 2.2 GB/2.2 GB 1.4 MB/s 0s
pulling fa8235e5b48f... 100% 1.1 KB
pulling 542b217f179c... 100% 148 B
pulling 8dde1baf1db0... 100% 78 B
pulling 23291dc44752... 100% 483 B
verifying sha256 digest
writing manifest
success
jetson@yahboom:~$
  
```

4. Use Phi-3

4.1. Run Phi-3

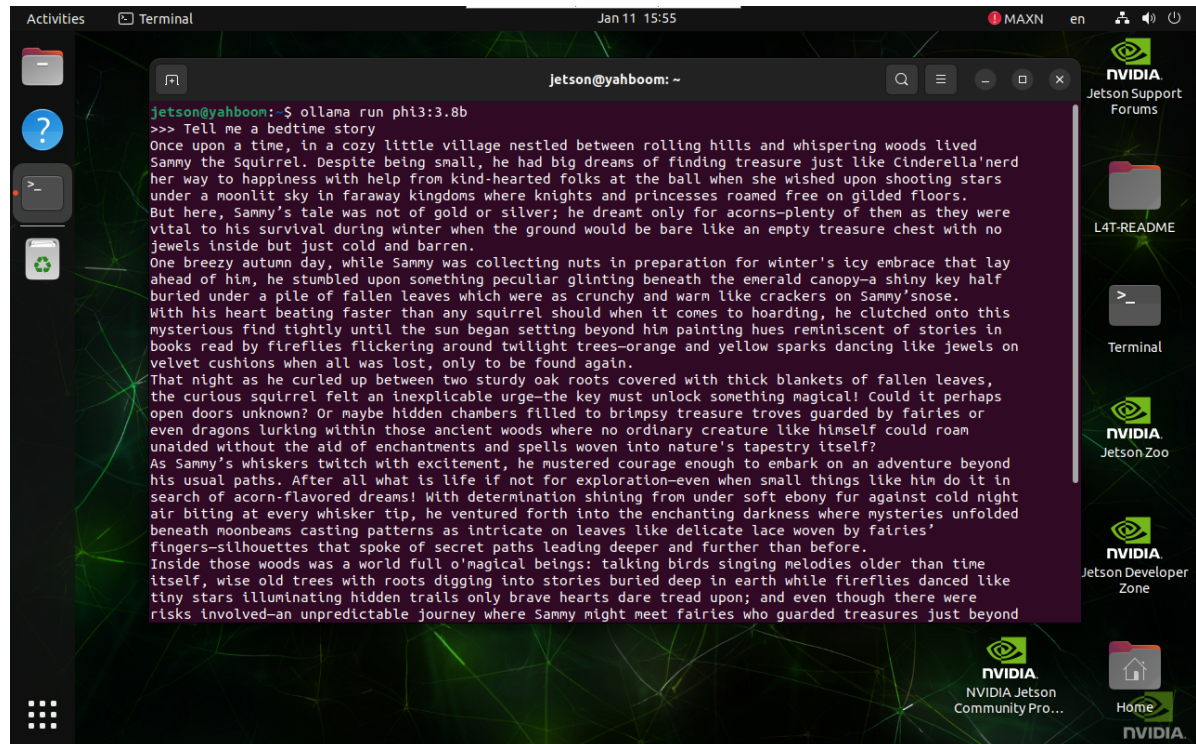
If the system does not have a running model, the system will automatically pull the Phi-3 3.8B model and run it:

```
ollama run phi3:3.8b
```

4.2. Have a conversation

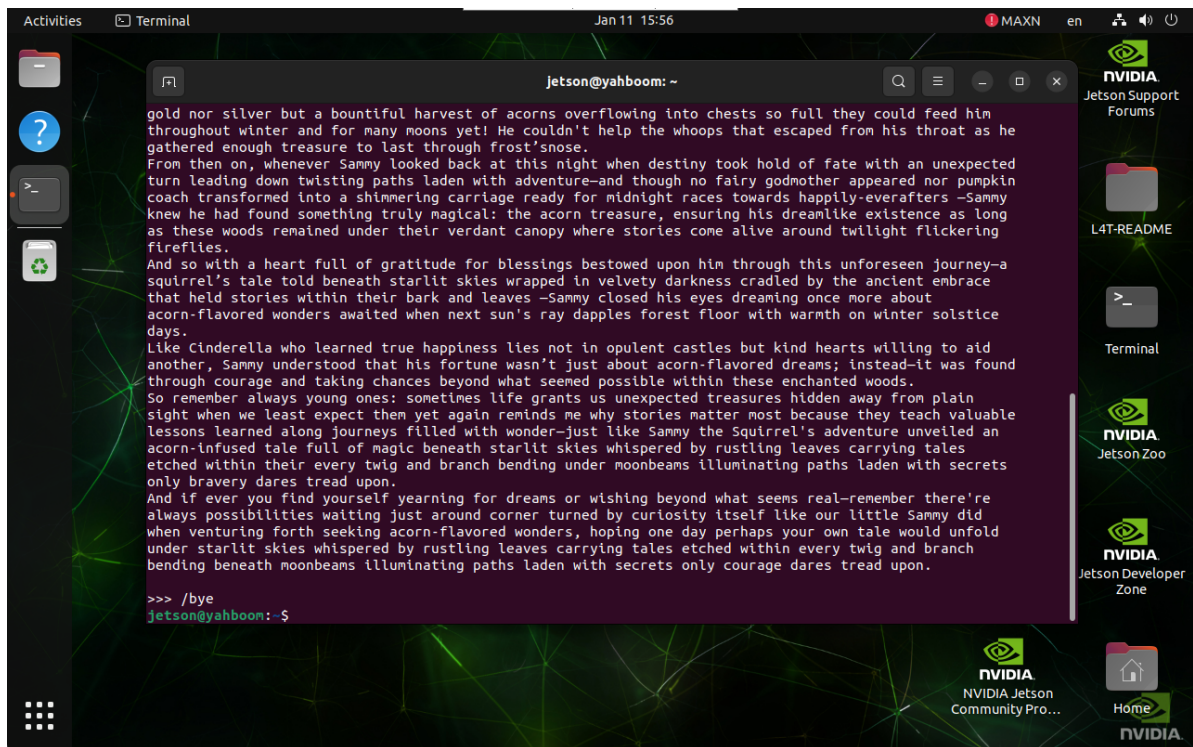
```
Tell me a bedtime story
```

The time to reply to the question depends on the hardware configuration, please wait patiently!



4.3. End the conversation

Use the `ctrl+d` shortcut key or `/bye` to end the conversation!



References

Ollama

Official website: <https://ollama.com/>

GitHub: <https://github.com/ollama/ollama>

Phi-3

Ollama corresponding model: <https://ollama.com/library/phi3>