Llama 3.2

> **Demo Environment**

Development Boards: Jetson Orin, RDK X5, Raspberry Pi 5 series motherboard

Note: Due to performance limitations, Jetson Orin Nano 4GB, Raspberry Pi 5 2GB/4GB, and RDK X5 4GB versions require a smaller parameter version to run.

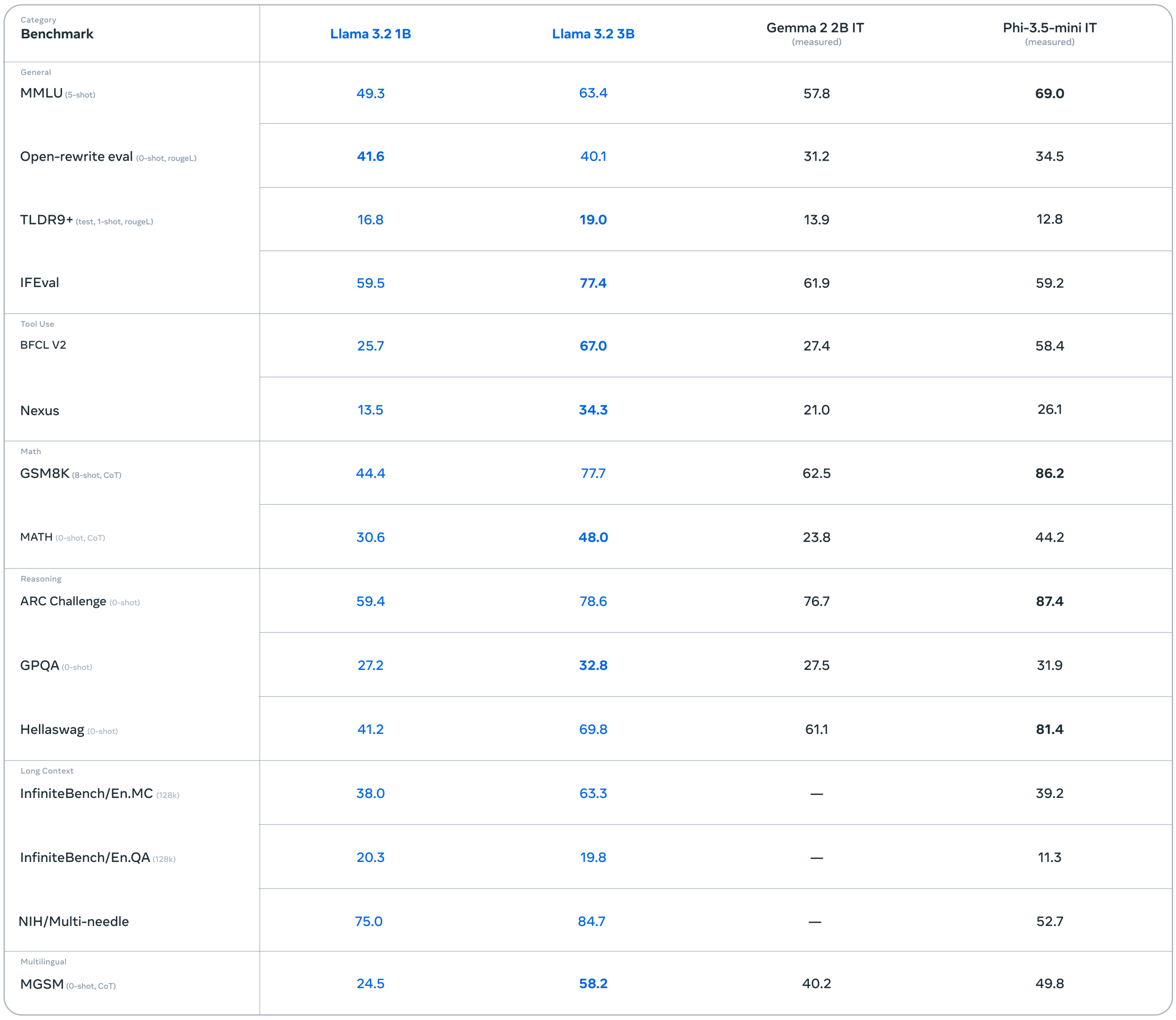
Meta Llama 3.2 is a series of advanced open-source large language models (LLMs) developed by the Meta AI department.

1. Model Size

Model	Size
llama3.2:1b	1.3GB
llama3.2:3b	2.0GB

2. Performance

Lightweight instruction-tuned benchmarks



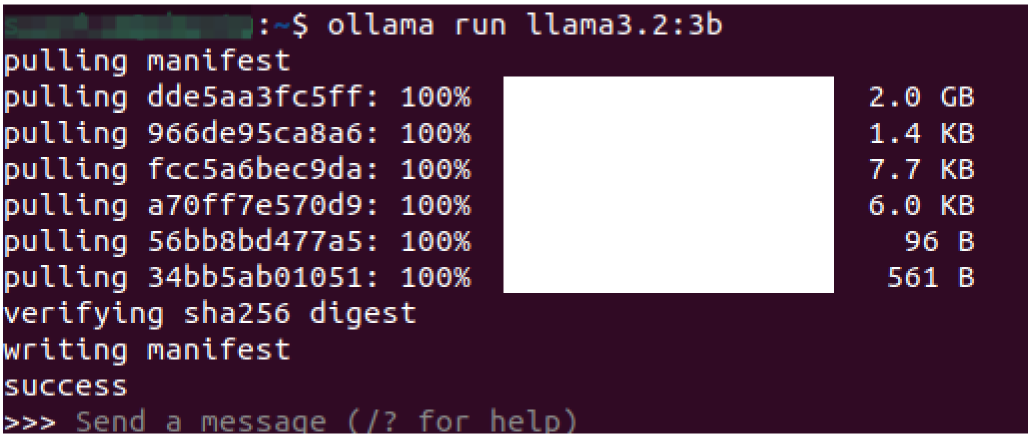
Category Benchmark	Llama 3.2 1B	Llama 3.2 3B	Gemma 2 2B IT (measured)	Phi-3.5-mini IT (measured)
General MMLU (5-shot)	49.3	63.4	57.8	**69.0**
Open-rewrite eval (0-shot, rougeL)	**41.6**	40.1	31.2	34.5
TLDR9+ (test, 1-shot, rougeL)	16.8	**19.0**	13.9	12.8
IFEval	59.5	**77.4**	61.9	59.2
Tool Use BFCL V2	25.7	**67.0**	27.4	58.4
Nexus	13.5	**34.3**	21.0	26.1
Math GSM8K (8-shot, CoT)	44.4	77.7	62.5	**86.2**
MATH (0-shot, CoT)	30.6	**48.0**	23.8	44.2
Reasoning ARC Challenge (0-shot)	59.4	78.6	76.7	**87.4**
GPQA (0-shot)	27.2	**32.8**	27.5	31.9
Hellaswag (0-shot)	41.2	69.8	61.1	**81.4**
Long Context InfiniteBench/En.MC (128k)	38.0	63.3	—	39.2
InfiniteBench/En.QA (128k)	20.3	19.8	—	11.3
NIH/Multi-needle	75.0	84.7	—	52.7
Multilingual MGSM (0-shot, CoT)	24.5	**58.2**	40.2	49.8

3. Using Llama 3.2

3.1 Running Llama 3.2

Use the `run` command to start running the model. If you haven't downloaded this model before, it will automatically pull the model from the Ollama model library:

```
ollama run llama3.2:3b
```

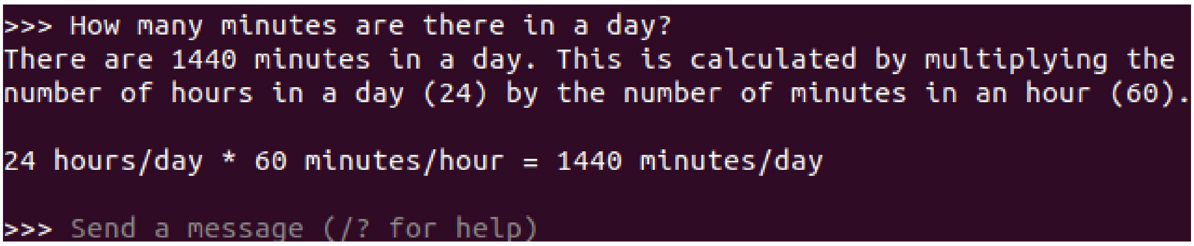


```
[redacted]:~$ ollama run llama3.2:3b
pulling manifest
pulling dde5aa3fc5ff: 100%                          2.0 GB
pulling 966de95ca8a6: 100%                          1.4 KB
pulling fcc5a6bec9da: 100%                          7.7 KB
pulling a70ff7e570d9: 100%                          6.0 KB
pulling 56bb8bd477a5: 100%                            96 B
pulling 34bb5ab01051: 100%                           561 B
verifying sha256 digest
writing manifest
success
>>> Send a message (/? for help)
```

3.2 Starting a Conversation

```
How many minutes are there in a day?
```

Response time depends on hardware configuration; please be patient!



```
>>> How many minutes are there in a day?
There are 1440 minutes in a day. This is calculated by multiplying the
number of hours in a day (24) by the number of minutes in an hour (60).

24 hours/day * 60 minutes/hour = 1440 minutes/day

>>> Send a message (/? for help)
```

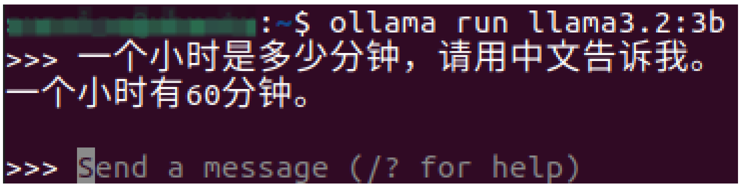
3.3 Ending a Conversation

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

3.4 Chinese Dialogue

For those without a Chinese input method, please refer to the Chinese input method switching tutorial.

Chinese Dialogue:



```
[redacted]:~$ ollama run llama3.2:3b
>>> 一个小时是多少分钟，请用中文告诉我。
一个小时有60分钟。

>>> Send a message (/? for help)
```

References

> **Ollama**

Official Website: https://ollama.com/

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

> **Llama 3.2**

Official Website: https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_2/

Ollama Corresponding Model: https://ollama.com/library/llama3.2