Phi4-mini

> **Demo Environment**

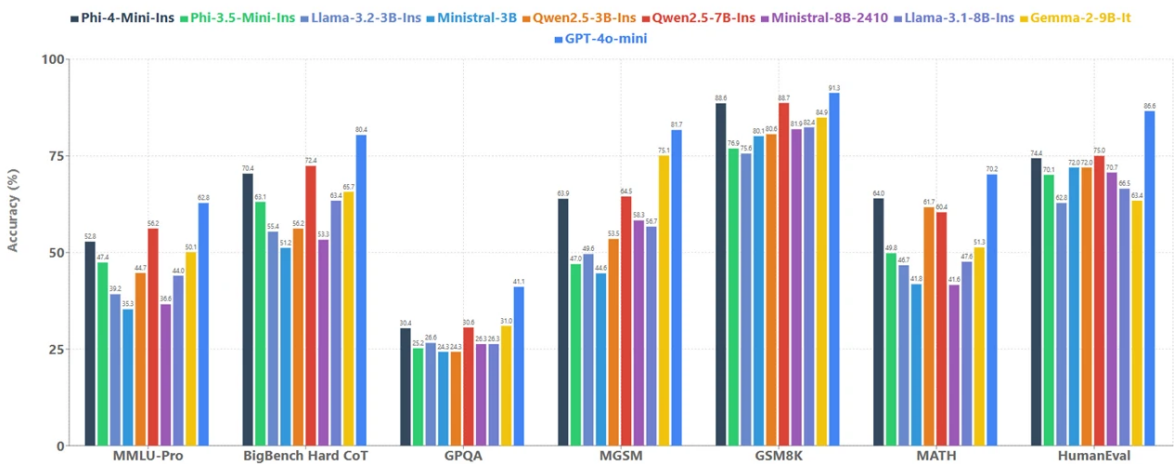**Development Boards**: Jetson Orin, RDK X5, Raspberry Pi 5 series motherboard

Note: Jetson Orin Nano 4GB and Raspberry Pi 5 2GB cannot run this model due to performance limitations.

Phi-4-mini-instruct is a lightweight, open model built on synthetic data and selected public websites, focusing on high-quality, inference-intensive data.

1. Model Size

Model	Size
phi4-mini:3.8b	2.5GB

2. Performance

3. Using Phi4-mini

3.1 Running Phi4-mini

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 Ollam model library:

```
ollama run phi4-mini:3.8b
```

```
                   :~$ ollama run phi4-mini:3.8b
pulling manifest
pulling 3c168af1dea0: 100%                          2.5 GB
pulling 813f53fdc6e5: 100%                           655 B
pulling fa8235e5b48f: 100%                           1.1 KB
pulling 8c2539a423c4: 100%                           411 B
verifying sha256 digest
writing manifest
success
>>> Send a message (/? for help)
```

3.2 Engaging in a Dialogue

```
How many minutes is a quarter of an hour?
```

The response time depends on your hardware configuration. Please be patient!

```
>>> How many minutes is a quarter of an hour?
A standard hour consists of 60 minutes. Therefore, one-quarter (1/4) of an
hour would be:

(1/4) * 60 = 15

So, there are 15 minutes in a quarter-hour.

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

3.3 Ending the Dialogue

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

3.4 Chinese Dialogue

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

Chinese Dialogue:

```
>>> 一个小时是多少分钟？请用中文告诉我
一小时等于 60 分钟。

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

References

> **Ollama**

Official Website: https://ollama.com/

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

> **Phi4-mini**

Ollama corresponding model: https://ollama.com/library/phi4-mini