Qwen3

Qwen3
    1. Model Size
    2. Performance
    3. Using Qwen3
        3.1 Running Qwen3
        3.2 Engaging in a Dialogue
        3.3 Ending the Dialogue
        3.4 Chinese Dialogue
    References

> **Demo Environment**

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

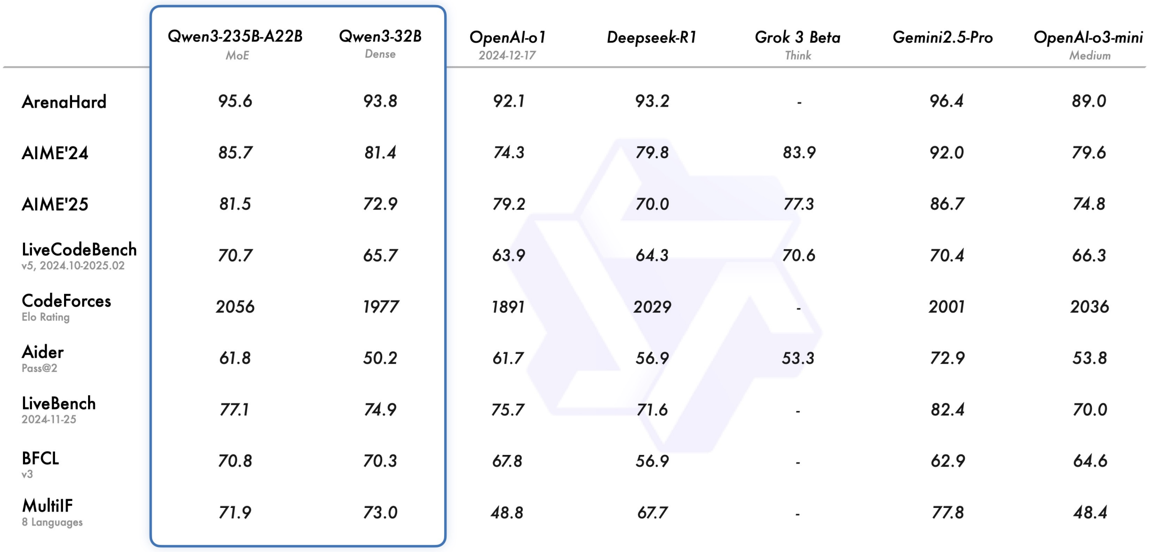
Note: Due to performance limitations, Jetson Orin Nano 4GB, Raspberry Pi 5 2GB/4GB, and RDK X5 4GB versions require a smaller parameter version to run.**

Qwen 3 is the latest generation of large language models in the Qwen series, providing a complete set of intensive and hybrid expert (MoE) models.

1. Model Size

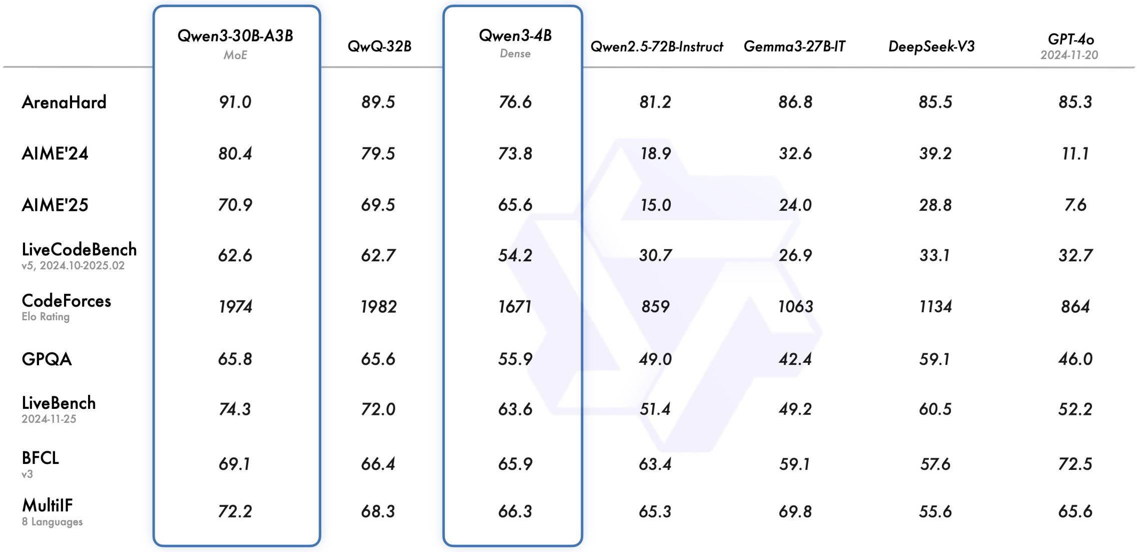
Model	Size
qwen3:0.6b	523MB
qwen3:1.7b	1.4GB
qwen3:4b	2.6GB
qwen3:8b	5.2GB

2. Performance



	Qwen3-235B-A22B MoE	Qwen3-32B Dense	OpenAI-o1 2024-12-17	Deepseek-R1	Grok 3 Beta Think	Gemini2.5-Pro	OpenAI-o3-mini Medium
ArenaHard	95.6	93.8	92.1	93.2	-	96.4	89.0
AIME'24	85.7	81.4	74.3	79.8	83.9	92.0	79.6
AIME'25	81.5	72.9	79.2	70.0	77.3	86.7	74.8
LiveCodeBench v5, 2024.10-2025.02	70.7	65.7	63.9	64.3	70.6	70.4	66.3
CodeForces Elo Rating	2056	1977	1891	2029	-	2001	2036
Aider Pass@2	61.8	50.2	61.7	56.9	53.3	72.9	53.8
LiveBench 2024-11-25	77.1	74.9	75.7	71.6	-	82.4	70.0
BFCL v3	70.8	70.3	67.8	56.9	-	62.9	64.6
MultiIF 8 Languages	71.9	73.0	48.8	67.7	-	77.8	48.4

1. AIME 24/25: We sample 64 times for each query and report the average of the accuracy. AIME'25 consists of Part I and Part II, with a total of 30 questions.
2. Aider: We didn't activate the think mode of Qwen3 to balance efficiency and effectiveness.
3. BFCL: The Qwen3 models are evaluated using the FC format, while the baseline models are assessed using the highest scores obtained from either the FC or prompt formats.



	Qwen3-30B-A3B MoE	QwQ-32B	Qwen3-4B Dense	Qwen2.5-72B-Instruct	Gemma3-27B-IT	DeepSeek-V3	GPT-4o 2024-11-20
ArenaHard	91.0	89.5	76.6	81.2	86.8	85.5	85.3
AIME'24	80.4	79.5	73.8	18.9	32.6	39.2	11.1
AIME'25	70.9	69.5	65.6	15.0	24.0	28.8	7.6
LiveCodeBench v5, 2024.10-2025.02	62.6	62.7	54.2	30.7	26.9	33.1	32.7
CodeForces Elo Rating	1974	1982	1671	859	1063	1134	864
GPQA	65.8	65.6	55.9	49.0	42.4	59.1	46.0
LiveBench 2024-11-25	74.3	72.0	63.6	51.4	49.2	60.5	52.2
BFCL v3	69.1	66.4	65.9	63.4	59.1	57.6	72.5
MultiIF 8 Languages	72.2	68.3	66.3	65.3	69.8	55.6	65.6

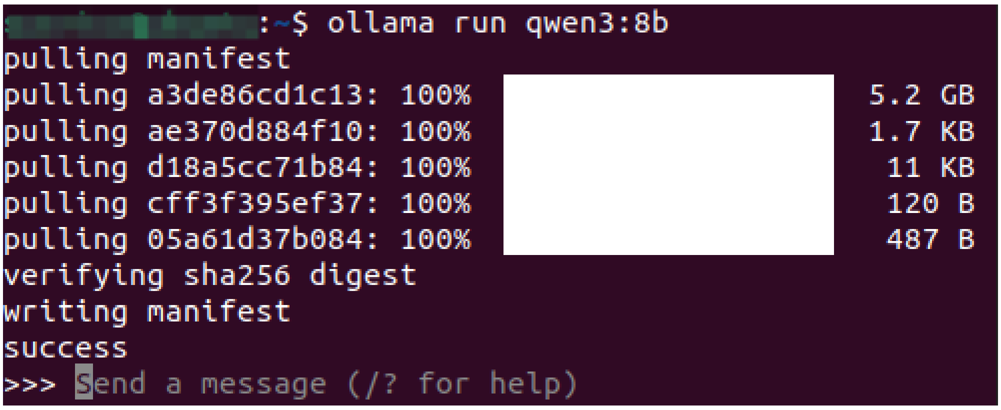
1. AIME 24/25: We sample 64 times for each query and report the average of the accuracy. AIME'25 consists of Part I and Part II, with a total of 30 questions.
2. Aider: We didn't activate the think mode of Qwen3 to balance efficiency and effectiveness.
3. BFCL: The Qwen3 models are evaluated using the FC format, while the baseline models are assessed using the highest scores obtained from either the FC or prompt formats.

3. Using Qwen3

3.1 Running Qwen3

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

```
ollama run qwen3:8b
```

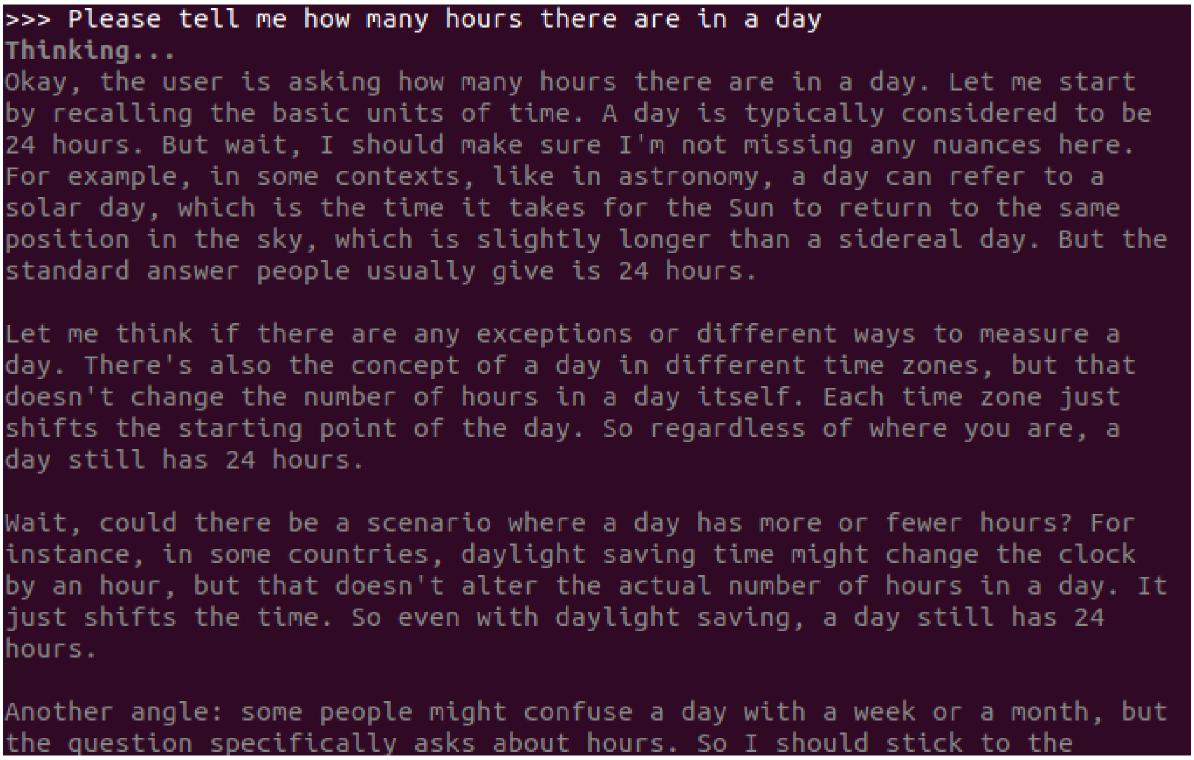


```
[redacted]:~$ ollama run qwen3:8b
pulling manifest
pulling a3de86cd1c13: 100%                        5.2 GB
pulling ae370d884f10: 100%                        1.7 KB
pulling d18a5cc71b84: 100%                         11 KB
pulling cff3f395ef37: 100%                         120 B
pulling 05a61d37b084: 100%                         487 B
verifying sha256 digest
writing manifest
success
>>> Send a message (/? for help)
```

3.2 Engaging in a Dialogue

```
Please tell me how many hours there are in a day
```

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



```
>>> Please tell me how many hours there are in a day
Thinking...
Okay, the user is asking how many hours there are in a day. Let me start
by recalling the basic units of time. A day is typically considered to be
24 hours. But wait, I should make sure I'm not missing any nuances here.
For example, in some contexts, like in astronomy, a day can refer to a
solar day, which is the time it takes for the Sun to return to the same
position in the sky, which is slightly longer than a sidereal day. But the
standard answer people usually give is 24 hours.

Let me think if there are any exceptions or different ways to measure a
day. There's also the concept of a day in different time zones, but that
doesn't change the number of hours in a day itself. Each time zone just
shifts the starting point of the day. So regardless of where you are, a
day still has 24 hours.

Wait, could there be a scenario where a day has more or fewer hours? For
instance, in some countries, daylight saving time might change the clock
by an hour, but that doesn't alter the actual number of hours in a day. It
just shifts the time. So even with daylight saving, a day still has 24
hours.

Another angle: some people might confuse a day with a week or a month, but
the question specifically asks about hours. So I should stick to the
```

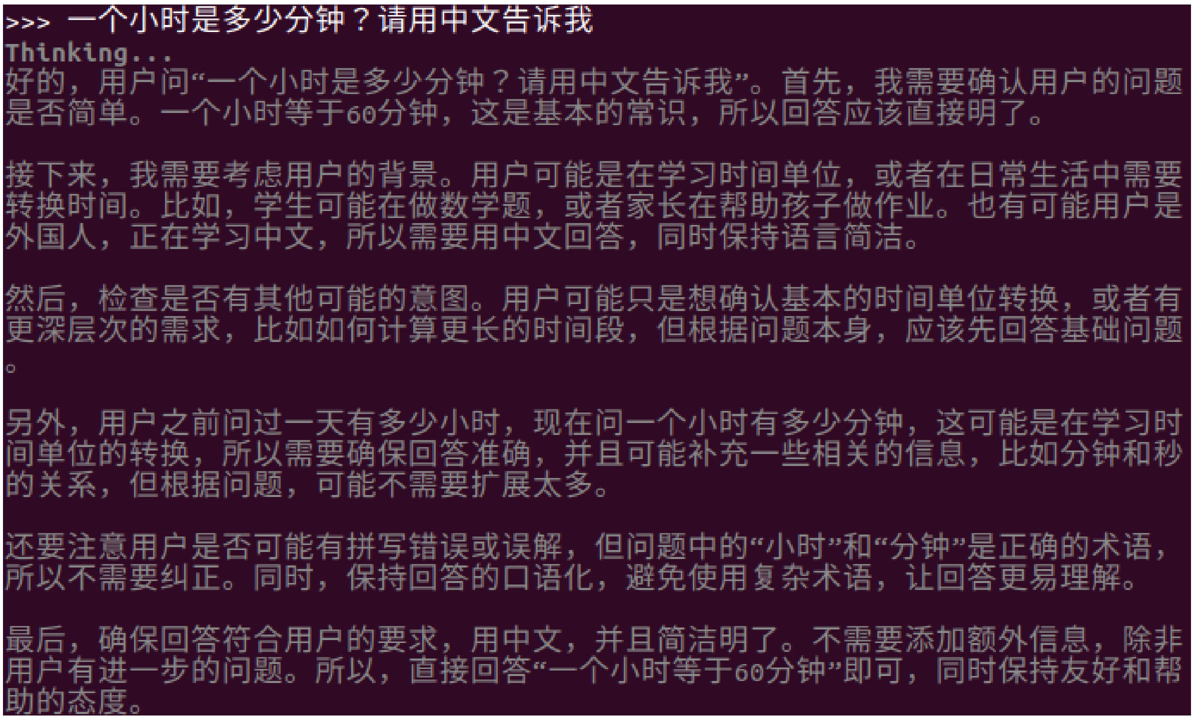
3.3 Ending the Dialogue

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
Thinking...
好的，用户问“一个小时是多少分钟？请用中文告诉我”。首先，我需要确认用户的问题
是否简单。一个小时等于60分钟，这是基本的常识，所以回答应该直接明了。

接下来，我需要考虑用户的背景。用户可能是在学习时间单位，或者在日常生活中需要
转换时间。比如，学生可能在做数学题，或者家长在帮助孩子做作业。也有可能用户是
外国人，正在学习中文，所以需要用中文回答，同时保持语言简洁。

然后，检查是否有其他可能的意图。用户可能只是想确认基本的时间单位转换，或者有
更深层次的需求，比如如何计算更长的时间段，但根据问题本身，应该先回答基础问题
。

另外，用户之前问过一天有多少小时，现在问一个小时有多少分钟，这可能是在学习时
间单位的转换，所以需要确保回答准确，并且可能补充一些相关的信息，比如分钟和秒
的关系，但根据问题，可能不需要扩展太多。

还要注意用户是否可能有拼写错误或误解，但问题中的“小时”和“分钟”是正确的术语，
所以不需要纠正。同时，保持回答的口语化，避免使用复杂术语，让回答更易理解。

最后，确保回答符合用户的要求，用中文，并且简洁明了。不需要添加额外信息，除非
用户有进一步的问题。所以，直接回答“一个小时等于60分钟”即可，同时保持友好和帮
助的态度。
```

References

> **Ollama**

Official Website: https://ollama.com/

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

> **Qwen3**

GitHub: https://github.com/QwenLM/Qwen3

Ollama Corresponding Model: https://ollama.com/library/qwen3