

Qwen3

Qwen3

- 1. Model Size
- 2. Performance
- 3. Using Qwen3
 - 3.1. Running Qwen3
 - 3.2. Starting a Conversation
 - 3.3. Ending the Conversation
- References

Demo Environment

Development Board: Jetson Nano

SD (TF) Card: 64GB

Recommended for models with 4B parameters or less

Qwen 3 is the latest generation of large-scale language models in the Qwen series, providing a comprehensive suite of dense and mixture-of-experts (MoE) models.

Model Storage Location

/usr/share/ollama/.ollama/models

1. Model Size

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

2. Performance

	Qwen3-235B-A22B <i>MoE</i>	Qwen3-32B <i>Dense</i>	OpenAI-o1 <i>2024-12-17</i>	Deepseek-R1	Grok 3 Beta <i>Think</i>	Gemini2.5-Pro	OpenAI-o3-mini <i>Medium</i>
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 <i>v5, 2024.10-2025.02</i>	70.7	65.7	63.9	64.3	70.6	70.4	66.3
CodeForces <i>Elo Rating</i>	2056	1977	1891	2029	-	2001	2036
Aider <i>Pass@2</i>	61.8	50.2	61.7	56.9	53.3	72.9	53.8
LiveBench <i>2024-11-25</i>	77.1	74.9	75.7	71.6	-	82.4	70.0
BFCL <i>v3</i>	70.8	70.3	67.8	56.9	-	62.9	64.6
MultilF <i>8 Languages</i>	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. AIME25 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 <i>MoE</i>	QwQ-32B	Qwen3-4B <i>Dense</i>	Qwen2.5-72B-Instruct	Gemma3-27B-IT	DeepSeek-V3	GPT-4o <i>2024-11-20</i>
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 <i>v5, 2024.10-2025.02</i>	62.6	62.7	54.2	30.7	26.9	33.1	32.7
CodeForces <i>Elo Rating</i>	1974	1982	1671	859	1063	1134	864
GPQA	65.8	65.6	55.9	49.0	42.4	59.1	46.0
LiveBench <i>2024-11-25</i>	74.3	72.0	63.6	51.4	49.2	60.5	52.2
BFCL <i>v3</i>	69.1	66.4	65.9	63.4	59.1	57.6	72.5
MultilF <i>8 Languages</i>	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. AIME25 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 the model is not already downloaded, it will automatically pull the model from the Ollama model library:

```
ollama run qwen3:1.7b
```

```
user@~$ 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. Starting a Conversation

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

Response time depends on your hardware configuration, so 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 Conversation

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

References

Ollama

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

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

Qwen3

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

Ollama Model: <https://ollama.com/library/qwen3>

