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 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

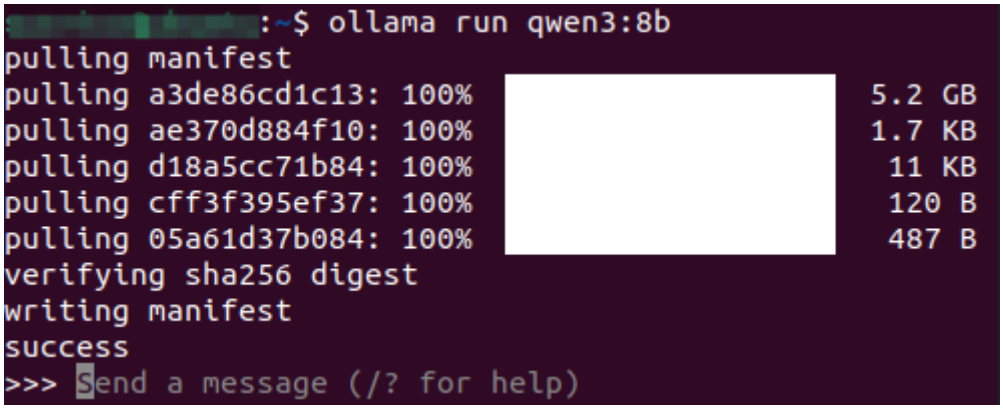
1. AIME 24/25: We sample 64 times for each query and report the average of the accuracy. AIME'25 consists of Part I and Part II, with a total of 30 questions.
2. Aider: We didn't activate the think mode of Qwen3 to balance efficiency and effectiveness.
3. BFCL: The Qwen3 models are evaluated using the 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 the model is not already downloaded, it will automatically pull the model from the Ollama model library:

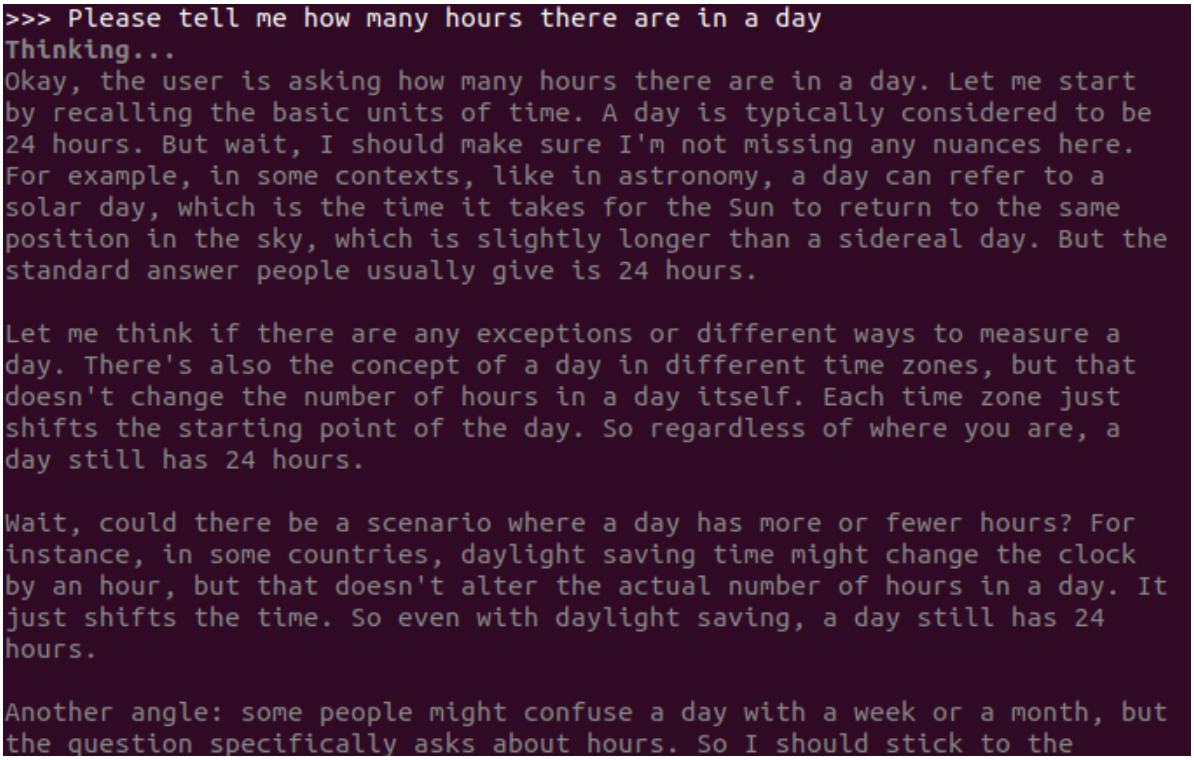
```
ollama run qwen3:1.7b
```



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!



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