

Server - ollama

- Installation and pulling models

On MacOS v26 (on MacBook Pro M5)

```
curl -fsSL https://ollama.com/install.sh | sh
```

In use

ollama pull qwen3.8:27b-mix

```
anton-w@antlmbp7 2026-09-25 - ollama new LLMs % ollama pull qwen3.8:27b-mix  
pulling manifest  
pulling model: 100% ████████████████████████████████████████████████████████████ 18 GB  
writing manifest  
success  
anton-w@antlmbp7 2026-09-25 - ollama new LLMs % ollama list
```

NAME	ID	SIZE	MODIFIED
qwen3.8:27b-mix	5642e97495e1	18 GB	2 minutes ago

ollama pull gwen3.6:35b-a3b

the official qwen3.6:35b-a3b tag is a 23GB Q4_K_M. That's tight on 32 GB, so if memory gets squeezed, pull a smaller community IQ4 XS instead (~17.7GB):

```
ollama pull hf.co/bartowski/Qwen_Qwen3.6-35B-A3B-GGUF:IQ4_XS
```

Small model: for autocomplete

```
ollama pull qwen3.5:9b
```

```
anton-w@ant1mbp7 ~ % ollama pull qwen3.5:9b
pulling manifest
pulling dec52a44569a: 100%
pulling 9371364b27a5: 100%
verifying sha256 digest
writing manifest
success
anton-w@ant1mbp7 ~ %
```



6.6 GB/6.6 GB	57 MB/s	0s
---------------	---------	----

Tested

- good, but too big, calculations are long (40'--4h), almost can not predict workflow pace (timing). Sometimes takes too much time for smaller tasks.

```
ollama pull qwen3-coder:30b
```