Daily AI Cooking

- Server - ollama
  - The Models Journal

Server - ollama

Server - ollama

The Models Journal

Ollama itself

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

on MacBook Pro M5 On MacOS v26

```
ollama pull qwen3.8:27b-mlx
```




```
ollama pull qwen3.6:35b-a3b
```

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




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

2026-10-05

Today I have decided to go with (relatively) smaller models, than earlier

```
ollama rm qwen3.6:35b-a3b    # each session was 1h-1h30 / (40'--4h)
ollama rm qwen3.8:27b-mlx    # each session was 1h-1h30 / (40'--4h)
ollama pull ministral-3:14b
ollama pull ornith:9b
ollama pull qwen3:14b
ollama rm qwen3.5:9b         # this was good



anton-w@ant1mbp7 ~ % ollama list
NAME                    ID              SIZE      MODIFIED
qwen3:14b               bdbd181c33f2    9.3 GB    2 hours ago    # switched after qwen3.5:9b
ornith:9b               a75697c14589    5.6 GB    2 hours ago    # good, different approached
ministral-3:14b         4760c35aeb9d    9.1 GB    2 hours ago
phi4:14b                ac896e5b8b34    9.1 GB    2 hours ago
gemma4:26b              001e5dafc3c7    18 GB     3 days ago
gpt-oss:20b             17052f91a42e    13 GB     6 days ago
devstral-small-2:24b    24277f07f62d    15 GB     6 days ago
```

on Lenovo ThinkCentre M93p

2026-10-04

Handy (speech-to-text **small** LLMs) for small desktop thin computer

English language

- Moonshine Streaming Small.
- Nemotron Speech Streaming EN
- Nemotron Streaming 3.5
- Parakeet Unified EN 0.6B

Russian language

- GigaAM v3 CTC
- GigaAM v3 E2E-CTC | Active)
- GigaAM v3 E2E-RNN-T

- GigaAM v3 RNN-T