Server - ollama

- Installation and pulling models

Installation and pulling models

On MacOS v26 (on MacBook Pro M5)

Ollama itself

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

Models

In use

Best quality: the Apple Silicon (MLX) build (~18 GB)

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

Fast: MoE daily driver (~23 GB)

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

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
```