Ollama Run LLMs locally

CONCEPTS

The CLI is a client: it talks to a local server on port 11434. Desktop and systemd installs start it for you.

Registry ollama.com
↓ pull
Local models ollama ls
↓ run
Server port 11434
↓ HTTP API
Clients CLI, curl, apps

Models stay in memory for 5 minutes after the last request, then unload. Change it with OLLAMA_KEEP_ALIVE.

Model names are name:tag; without a tag you get latest. Pick a tag to choose a size or quantization.

RUN AND CHAT

ollama run *<model>*
Start an interactive chat

ollama run *<model>* "*<prompt>*"
Answer one prompt and exit

cat *<file>* **| ollama run** *<model>* "*<prompt>*"
Feed a file as context

ollama run *<model>* **--verbose**
Show timings and tokens/s

ollama run *<model>* **--format json**
Force JSON output

ollama run *<model>* **--think=false**
Turn thinking off or set its effort

ollama run *<model>* **--keepalive** *<dur>*
Keep the model loaded, e.g. 30m

"""
Start and end multi-line input

MANAGE MODELS

ollama pull *<model>*
Download or update a model

ollama ls
List downloaded models

ollama ps
List loaded models

MANAGE MODELS (cont.)

ollama stop *<model>*
Unload a running model

ollama rm *<model>*
Remove a model

ollama show *<model>*
Show details of a model

ollama show *<model>* **--modelfile**
Print its Modelfile

ollama cp *<source>* *<new-name>*
Copy a model under a new name

ollama create *<name>* **-f** *<Modelfile>*
Build a model from a Modelfile

IN-CHAT COMMANDS

/? / /help
Show help for slash commands

/bye
Leave the chat

/clear
Clear the session context

/set system *<text>*
Set the system message

/set parameter *<name>* *<value>*
Change a parameter, e.g. temperature

/set verbose / /set quiet
Show or hide stats after replies

/set think / /set nothink
Turn thinking on or off

/show info
Show model details

/save *<name>*
Save this session as a model

/load *<name>*
Load a saved session or a model

MODELFILE

Custom model

```
FROM llama3.2
PARAMETER temperature 0.7
PARAMETER num_ctx 8192
SYSTEM """You are a concise assistant."""
```

FROM *<model or path>*
Base model, GGUF file or weights dir

MODELFILE (cont.)

PARAMETER *<name>* *<value>*
Set a runtime parameter

SYSTEM """*<text>***"""**
Set the system message

PARAMETER num_ctx *<n>*
Context window in tokens

SERVER AND ENVIRONMENT

ollama serve
Run the server in the foreground

OLLAMA_HOST=0.0.0.0:11434
Listen on all interfaces

OLLAMA_HOST=*<host:port>*
ollama ls
Point the CLI at another server

OLLAMA_MODELS=*<dir>*
Where models are stored

OLLAMA_KEEP_ALIVE=*<dur>*
How long idle models stay loaded

OLLAMA_CONTEXT_LENGTH=*<n>*
Default context window in tokens

systemctl edit ollama.service
Set variables for the Linux service

REST API

http://localhost:11434/api
Base URL of the local API

POST /api/chat
Chat with message history

POST /api/generate
Complete a single prompt

http://localhost:11434/v1
OpenAI-compatible base URL

Responses stream by default; send "stream": false for one JSON reply.

RECIPES

Ask over HTTP

```
curl http://localhost:11434/api/chat -d '{
  "model": "<model>",
  "messages": [{"role": "user", "content": "Hello"}],
  "stream": false
}' | jq -r .message.content
```