Here's a concise cheat sheet for common **Ollama** commands, assuming you're using Ollama to run and manage local large language models. These commands are executed in a terminal or command-line interface where Ollama is installed.

#### **Basic Ollama Commands**

Command	Description	
ollamaversion	Check the installed version of Ollama.	
ollama help	Display help information for Ollama commands.	
ollama list	List all available models installed locally.	
ollama pull <model_name></model_name>	Download a model from the Ollama registry (e.g., ollama pull llama3).	
ollama run <model_name></model_name>	Run a model interactively (e.g., ollama run llama3).	
ollama ps	Show currently running models and their processes.	
ollama stop <model_name></model_name>	Stop a running model.	
ollama rm <model_name></model_name>	Remove a model from your local system.	
<pre>ollama cp <source_model> <new_model></new_model></source_model></pre>	Copy a model to create a new one with a different name.	

### **Model Management**

Command	Description
<pre>ollama create <model_name> - f <file></file></model_name></pre>	Create a custom model from a Modelfile (e.g., ollama create mymodel -f Modelfile).
ollama show <model_name></model_name>	Display details about a specific model (e.g., configuration, parameters).
ollama push <model_name></model_name>	Push a local model to the Ollama registry (if supported).

## **Running Models**

Command	Description
<pre>ollama run <model_name> " <pre>cprompt&gt;"</pre></model_name></pre>	Run a model with a specific prompt non-interactively (e.g., ollama run llama3 "Hello!").
ollama run <model_name> verbose</model_name>	Run a model with verbose output for debugging.
<pre>ollama run <model_name> port <port></port></model_name></pre>	Specify a custom port for the model server (default is 11434).

## **API Interaction**

Ollama provides a REST API for programmatic access. Common endpoints:

Command	Description
curl http://localhost:11434/api/tags	List available models via the API.
<pre>curl -X POST http://localhost:11434/api/generate -d '{"model": "<model_name>", "prompt": "<pre>prompt&gt;"}'</pre></model_name></pre>	Generate a response from a model via API.
curl http://localhost:11434/api/ps	Check running models via

# **Configuration and Setup**

- \* Default Port: Ollama runs on http://localhost:11434 by default.
- \* Environment Variables:
  - \*OLLAMA\_HOST: Change the default host/port (e.g., export OLLAMA\_HOST=0.0.0.0:8080).
  - \*OLLAMA\_MODELS: Set the directory for storing models (e.g., export OLLAMA\_MODELS=/path/to/models).
- \* Modelfile: A file to define custom models (e.g., specify base model, parameters, or system prompts).

#### **Examples**

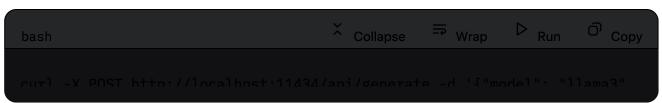
<sup>1</sup>·Pull and Run a Model:



<sup>2</sup> Create a Custom Model:



<sup>3</sup> API Request:



#### **Notes**

- \* Model Names: Use specific model tags (e.g., llama3:8b, mistral:latest) to specify versions or sizes.
- \* System Requirements: Ensure sufficient CPU/GPU and memory for the model you're running.
- Documentation: For advanced usage, check the official Ollama documentation at <a href="https://ollama.com">https://ollama.com</a> or GitHub.

If you need more details or specific examples, let me know!