Running Ollama on Google Colab

Quick setup guide for running local LLM inference using Ollama in Google Colab notebooks.

Step 1: Install and Run xterm

Install the colab-xterm package:

```python
!pip install colab-xterm
```

Load the xterm extension:

```python
%load_ext colabxterm
```

Launch xterm terminal:

```python
%xterm
```

Step 2: Install and Run Ollama (in xterm)

Inside the xterm terminal, run the following commands:

Install Ollama:

```bash
curl https://ollama.ai/install.sh | sh
```

Start Ollama server in background:

```bash
ollama serve &
```

Step 3: Download and Run a Model

Example - Run Llama 3.2 1B model:

```bash
ollama run llama3.2:1b
```

Other Popular Models

- `ollama run llama3.2:3b` - Larger Llama model
- `ollama run mistral` - Mistral 7B
- `ollama run codellama` - Code-specialized model
- Visit ollama.com/library for full model list

Quick Reference Commands

Command	Description
`ollama list`	List installed models
`ollama rm model_name`	Remove a model
`ollama pull model_name`	Pull model without running
`ollama run model_name`	Download and run a model

Notes

- All commands in Step 2 and 3 are executed inside the xterm terminal
- The `&` symbol runs Ollama server in the background
- Models are downloaded on first run and cached for subsequent use
- Free Colab instances have limited GPU/RAM - choose smaller models if needed

---

Resources

- Ollama Documentation
- Ollama Model Library

- Google Colab