# **Running Small Language Models (SLMs) with Ollama on Windows**

This guide walks you through everything from installing **Ollama** to running **Small Language Models** like **Phi-2, Mistral, LLaMA 2, and Gemma**, and optionally integrating a **simple web-based UI** using **Streamlit**.

## What Are SLMs?

**Small Language Models (SLMs)** are compact versions of large AI models that run efficiently on personal devices. They are fast, lightweight, and ideal for use cases where privacy, cost, and performance matter.

Common SLMs:

* **Phi-2** – Efficient reasoning and education
* **Mistral** – Versatile general-purpose model
* **LLaMA 2** – Balanced performance across tasks
* **Gemma** – Google's model for safe

## Step-by-Step Setup Guide

### Step 1: Install Ollama

1. Go to: <https://ollama.com>
2. Click **Download for Windows**
3. Run the .exe installer
4. Follow the on-screen instructions
5. After install, Ollama runs silently in the background

### Step 2: Pull a Model (First Time Only)

1. Open **Command Prompt** (Win + R, type cmd, Enter)
2. Type one of the following to download a model:

ollama pull phi # Pull Phi-2 (lightweight reasoning)

ollama pull mistral # Pull Mistral (fast, versatile)

ollama pull llama2 # Pull Meta's LLaMA 2

ollama pull gemma # Pull Google's Gemma

### Step 3: Run the Model

Once downloaded, run the model in chat mode:

ollama run phi

You’ll see a prompt like:

>>>

Now type your questions, for example:

>>> What is OEE in manufacturing?

>>> Explain predictive maintenance.

To exit the chat:  
Press Ctrl + C

**Add a Simple Web UI using Streamlit**

**Step 4.1: Install Python & Streamlit**

If Python is not installed:

1. Go to [https://python.org](https://python.org/downloads)
2. Download and install Python for Windows
3. During setup, **check**: “Add Python to PATH”

Then install Streamlit:

pip install streamlit

**Step 4.2: Create the UI Script**

Create a file named ollama\_ui.py and paste this code:

import streamlit as st

import requests

st.title("Chat with Ollama Model")

prompt = st.text\_area("Enter your prompt:")

if st.button("Send"):

response = requests.post(

"http://localhost:11434/api/generate",

json={

"model": "phi", # Change to mistral, llama2, or gemma as needed

"prompt": prompt,

"stream": False

}

)

result = response.json()

st.text\_area("Response:", value=result["response"], height=200)

**Step 4.3: Run the UI**

In the terminal, go to the folder containing ollama\_ui.py and run:

streamlit run ollama\_ui.py