How to Run an LLM Model Locally on Windows Using Ollama

Ollama is a lightweight tool to run large language models (LLMs) locally without complex setup. Follow these steps:

✅ 1. Install Ollama on Windows

1. Download **Ollama for Windows** from the official site: 👉 https://ollama.com/download

2. Run the installer and complete the setup.

3. After installation, open **Command Prompt** or **PowerShell** and verify:

```
ollama --version
```

✅ 2. Pull a Model

Choose a model and pull it to your machine. For example:

```
ollama pull llama2
```

Other popular models:

- `mistral`
- `codellama`
- `phi`

✅ 3. Run the Model

Start the model in interactive mode:

```
ollama run llama2
```

Run a single prompt:

```
ollama run llama2 "Explain quantum computing in simple terms"
```

✅ 4. Run Ollama API (Local Server)

Ollama exposes a local API on **http://localhost:11434** by default.

Example using **PowerShell (Invoke-RestMethod)**:

```
Invoke-RestMethod -Uri http://localhost:11434/api/generate -Method Post -Body '{"model": "llama2", "prompt": "Write a haiku about programming."}' -ContentType "application/json"
```

✅ 5. Integrate with Your App (Python Example)

```
import requests

url = "http://localhost:11434/api/generate"
data = {
    "model": "llama2",
    "prompt": "Write a Python function to reverse a string"
}
response = requests.post(url, json=data, stream=True)

for chunk in response.iter_lines():
    print(chunk.decode())
```

✅ 6. Hardware Requirements

- **CPU Mode:** Works but slower.
- **GPU Acceleration:** Supports NVIDIA GPUs.
- **RAM:** At least **8 GB**, recommended **16–32 GB** for larger models.

✅ 7. Custom Models (Optional)

Create a `Modelfile`:

```
FROM llama2
SYSTEM "You are an AI assistant specialized in coding."
```

Build your custom model:

```
ollama create my-model -f Modelfile
```

Run it:

```
ollama run my-model
```

✅ Next Steps

Do you want:

- **A)** GPU optimization guide for Windows?
- **B)** Full example of Ollama + .NET + Angular integration for a chat app?
- **C)** LangChain integration on Windows?