GPT-OSS Local Setup with Ollama
================================

Step 1: Install Ollama
----

1. Download and install Ollama from:
https://ollama.com/download

2. Verify installation:
```
ollama --version
```

Step 2: Pull the GPT-OSS Model

If GPT-OSS exists in Ollama's registry:
```
ollama pull gpt-oss
```

If not, create a custom model:
1. Create a file named "Modelfile" with content:
```
FROM hf://nvidia/GPT-OSS
```

2. Build the model locally:
```
ollama create gpt-oss -f Modelfile
```

Step 3: Run GPT-OSS from Terminal

```
ollama run gpt-oss
```

This starts an interactive chat in your terminal.

Step 4: Run GPT-OSS from Python
----

1. Install the Ollama Python client:
```
pip install ollama
```

2. Example code:
```
import ollama

response = ollama.chat(model='gpt-oss', messages=[
{'role': 'system', 'content': 'You are a helpful assistant.'},
{'role': 'user', 'content': 'Explain black holes in simple terms.'}
])

print(response['message']['content'])
```

Step 5: One-Click Start Script (Optional)

Create a file named "start.bat" with content:

```
@echo off
echo Starting GPT-OSS locally...
ollama run gpt-oss
pause
```

Double-click to start GPT-OSS instantly.

Notes:

- Make sure your system has enough RAM and GPU VRAM for large models.
- Hugging Face model link: https://huggingface.co/nvidia/GPT-OSS
- For advanced usage, see Ollama docs: https://github.com/ollama/ollama