🧑‍🏫 INSTRUCTOR SCRATCHER: Deploying Ollama on Self-Hosted VPS + LLM API Setup

✅ OVERVIEW

Goal: Students will learn to:

1. Deploy a VPS on Hostinger with Ubuntu 24.04
2. Install and configure Ollama
3. Download a free open-source LLM (e.g. Mistral)
4. Open Ollama API to external access using static ports
5. Use `socat` to tunnel internal API externally
6. Test the Ollama API with a prompt

🧱 STEP 0: VPS & OS SETUP

Step	Description	Command / Notes
0.1	Buy VPS from Hostinger	Use plan: **KVM 2 VPS**
0.2	Choose OS	Ubuntu 24.04 with **Ollama**
0.3	Open browser-based terminal from Hostinger dashboard	Useful for beginners. SSH also possible.

⚙️ STEP 1: UPDATE & INSTALL OLLAMA

Step	Description	Command / Notes
1.1	Update & upgrade system packages (Optional)	`sudo apt update && sudo apt upgrade -y`
1.2	Install Ollama **(SKIP)**	`curl -fsSL https://ollama.com/install.sh

		(only for new ubuntu not in prepaid version)
1.3	Verify installation	`ollama --version`
1.4	Update Ollama to latest version	`curl -fsSL https://ollama.com/install.sh | sh`

🌍 STEP 2: Define Static Ports & Enable External Access (Port Scratcher)

This step ensures that all your services (like Ollama API, Open WebUI, etc.) are assigned fixed, memorable ports — and that those ports are opened on the server to allow access.

🟢 2.1 Port Scratcher Table (Document for Every Project)

Service	Port	Protocol	Notes
Ollama (API)	11434	TCP	Main API used by AI Agents
Open WebUI	3000	TCP	Optional interface for interacting with Ollama
n8n (Remote)	5678	TCP	If running from Render or another server
SSH	22	TCP	To connect to the server
(Other services)	…	…	Add based on your own project needs

🟢 2.2 Open the Necessary Ports on Ubuntu

Make sure your firewall (UFW) allows traffic to these ports. Run the following commands on your VPS:

bash

```
sudo ufw allow 11434    # Ollama API
sudo ufw allow 3000     # Web UI (optional)
sudo ufw allow 22       # SSH (for remote access)
sudo ufw reload
sudo ufw status
```

> 🔒 **Security Note**: If you are not using the WebUI or SSH on public networks, limit access using IP whitelisting or VPN later in production.

🧠 STEP 3: DOWNLOAD LLM MODEL

Step	Description	Command / Notes
3.1	Pull the open-source model	`ollama pull mistral`
3.2	Confirm model exists	`ollama list`

Other : ollama pull phi3:mini OR ollama pull openhermes

🟢 4. 🚦 الخطوة الجاية: تأكيد فتح Ollama للعالم

1️⃣ شوف Ollama سامع على إيه فعليًا:

```
sudo netstat -tuln | grep 11434
ss -tuln | grep 11434
```

هل النتيجة:

- `127.0.0.1:11434` ← سامع محلي فقط (لازم نعدل)
- `0.0.0.0:11434` أو `:::11434` ← سامع على كل الشبكة (كده جاهز للعالم)

---

5 . عدل ملف خدمة Ollama ونستخدم البارامیتر الصحيح على حسب إصدارك.

🔁 Expose Ollama's API with Socat (Persistent System Service)

By default, Ollama’s internal API (`localhost:11434`) is not exposed to the outside world. We use **socat** to forward this local port to a public one (`0.0.0.0:11435`) so that external services and agents can access it.

✅ 5.1 Install `socat` (if not already installed)

```
sudo apt update
sudo apt install socat -y
```

✅ 5.2 Create a systemd service to keep Socat running automatically

```
sudo nano /etc/systemd/system/socat-ollama.service
```

Paste the following configuration

```
[Unit]
Description=Socat proxy to expose Ollama on 0.0.0.0

[Service]
ExecStart=/usr/bin/socat TCP-LISTEN:11435,reuseaddr,fork TCP:127.0.0.1:11434
Restart=always

[Install]
WantedBy=multi-user.target
```

✅ 5.3 Enable and start the Socat service

```
sudo systemctl daemon-reload
sudo systemctl enable socat-ollama
sudo systemctl start socat-ollama
sudo systemctl status socat-ollama
```

💡 You can now access your Ollama model externally via:

```
http://<your-server-ip>:11435
```

🚦 الخطوات اللي تعملها دلوقتي:

1️⃣ اختبر من السيرفر نفسه:

```
curl http://127.0.0.1:11434
```

لازم يطلع:

```
 Ollama is running
```

2️⃣ اختبر من السيرفر على البورت الخارجي الجديد:

```
curl http://localhost:11435
```

4️⃣ الآن استخدم خارجي ربط لأي:

http://69.62.118.174:11435

اختبار سريع لاستدعاء LLM على سيرفرك: check model name

```
curl -X POST http://69.62.118.174:11435/api/generate \

  -H "Content-Type: application/json" \

  -d '{"model":"mistral","prompt":"Hello from Ahmed, are you working?"}'
```
