Utilisation de l'API

? 1. Interroger l’API locale d’Ollama

Par défaut, l’API écoute sur :

```
http://localhost:11434
```

Exemple de requête API :

```
curl http://localhost:11434/api/generate \
  -d '{
        "model": "llama3",
        "prompt": "Bonjour, comment vas-tu ?"
      }'
```

? 2. Utilisation de l’API Ollama depuis Docker

Si vous exécutez des applications dans un **conteneur Docker** qui doivent appeler l’API Ollama située sur l’hôte, utilisez l’adresse suivante :

```
http://host.docker.internal:11434
```

Exemple depuis un conteneur :

```
curl http://host.docker.internal:11434/api/generate \
  -d '{"model": "llama3", "prompt": "Test depuis docker"}'
```

Remarque importante

- `host.docker.internal` fonctionne **automatiquement sur Docker Desktop (Windows / macOS)**.
- Sur Linux, Docker ne fournit **pas nativement** ce DNS.\ S'il n'existe pas, vous pouvez forcer la résolution en ajoutant dans votre commande `docker run` :

```
docker run --add-host=host.docker.internal:host-gateway ...
```

Ainsi, l'accès depuis Docker fonctionnera comme prévu.

? 3. Exemple simple depuis un script Python

```
import requests

payload = {
    "model": "llama3",
    "prompt": "Bonjour depuis Python !"
}

r = requests.post("http://localhost:11434/api/generate", json=payload)
print(r.text)
```

Revision #3
Created 2026-01-23 12:45:51 UTC by Maxime
Updated 2026-01-23 12:49:02 UTC by Maxime