AI Inference Gateway and Model Router by Code Creator

Customer Instructions - Ubuntu 24.04

This product provides a private CPU based AI inference gateway with Open WebUI, Ollama, LiteLLM, PostgreSQL, Redis, and Nginx. It includes a browser based AI chat interface and an OpenAI compatible API endpoint for local model access.

1. Recommended AWS Instance

Item	Recommendation
Minimum testing	t3.xlarge or larger
Recommended general use	m6i.xlarge, m7i.xlarge, m6i.2xlarge, or m7i.2xlarge
Storage	60 GB gp3 or larger
Operating system user	ubuntu

Smaller instances may start the web interface but can be slow during model startup. Larger optional models require more memory.

2. Required Security Group Ports

Type	Protocol	Port	Purpose
SSH	TCP	22	Administrator SSH access
HTTP	TCP	80	Open WebUI, help page, and API proxy

Restrict SSH port 22 to your own IP address whenever possible. Port 80 must be open to users who need browser or API access.

3. Connect by SSH

Connect to your instance using your Amazon private key and the **ubuntu** user.

```
ssh -i /path/to/your-key.pem ubuntu@PUBLIC_IP
```

4. First Login Instructions

After connecting by SSH, view the first login instructions:

```
cat /home/ubuntu/FIRST_LOGIN.txt
```

The file shows the current public IP based URLs for Open WebUI, the Code Creator help page, API information, and the OpenAI compatible API endpoint.

5. Application URLs

Service	URL Pattern
Open WebUI	http://PUBLIC_IP/
Code Creator Help Page	http://PUBLIC_IP/codecreator.html
API Information	http://PUBLIC_IP/api-info.html
OpenAI Compatible API Base	http://PUBLIC_IP/v1

On the first launch of a brand new instance, **Open WebUI may require 2 to 5 minutes to complete initial startup and model initialization.** If you receive a temporary 502 Bad Gateway message, wait a few minutes and refresh the page. This is expected on first startup.

6. Create the First Open WebUI Administrator

Open the Open WebUI URL in your browser:

http://PUBLIC_IP/

Create the first administrator account when prompted. Store the username and password securely. The first user created becomes the administrator for the Open WebUI interface.

7. Helpful Commands

Purpose	Command
Show URLs	codecreator-ai-url
Check status	codecreator-ai-status
List installed models	codecreator-ai-models
Install another Ollama model	codecreator-ai-install-model MODEL_NAME
View logs	codecreator-ai-logs
Restart the stack	codecreator-ai-restart

codecreator-ai-url

codecreator-ai-status

codecreator-ai-models

codecreator-ai-install-model MODEL_NAME

codecreator-ai-logs

codecreator-ai-restart

8. Default Model

The default CPU friendly model is llama3.2:3b. This model is included to make first launch easier on non GPU instances.

codecreator-ai-models

9. Optional Larger Models

You can install additional Ollama compatible models. Larger models require more memory and may not work well on smaller instances.

codecreator-ai-install-model llama3.1:8b

10. API Usage

The product exposes an OpenAI compatible API through LiteLLM at:

http://PUBLIC_IP/v1

Example chat completions request:

```
curl http://PUBLIC_IP/v1/chat/completions \
  -H "Authorization: Bearer YOUR_API_KEY" \
  -H "Content-Type: application/json" \
  -d '{"model":"llama3.2-3b","messages":[{"role":"user","content":"Say
hello"}],"max_tokens":40}'
```

Before production use, rotate placeholder API keys and store secrets securely.

11. Service Management

The Docker stack is configured to start automatically at boot using systemd. To manually restart the application stack:

codecreator-ai-restart

To inspect the status:

codecreator-ai-status

12. Documentation Resources

Open WebUI Documentation: https://docs.openwebui.com/

LiteLLM Documentation: https://docs.litellm.ai/

Ollama Documentation: https://docs.ollama.com/

Ollama Model Library: https://ollama.com/library

13. Troubleshooting

Temporary 502 Bad Gateway: **wait 2 to 5 minutes after first launch** and refresh the page. Open WebUI initializes its database and local cache during first startup.

Slow model response: use a larger instance size or install a smaller model.

Cannot reach the web interface: confirm port 80 is open in the EC2 security group and run:

codecreator-ai-status