Private AI Chat Sever GPT

Usage instructions:

1. Launch the product via 1-click from AWS Marketplace. **Wait** until the instance status changes to 'Running' and passes all health checks. Then, connect to your instance using your Amazon private key and the '**ubuntu**' user."

To update software, use: **sudo apt update && sudo apt upgrade -y**

2. Note: **Warm-up:** after first boot, the model may take **2–3 minutes** to finish loading. That's normal. Be Patient before running the test:

Run a smoke test:

```
sudo bash /opt/agent-server/smoke-test.sh
```

- You should see all 4 containers (nginx, open-webui, ollama, qdrant) and "GPU OK".

3. **Open the site**

In your browser go to:

- http://Your_Instance_Public_IP/

You'll land on the **Get started** page and be prompted to **create the first admin account**. (This is expected on fresh launches.)

Inside Open WebUI:

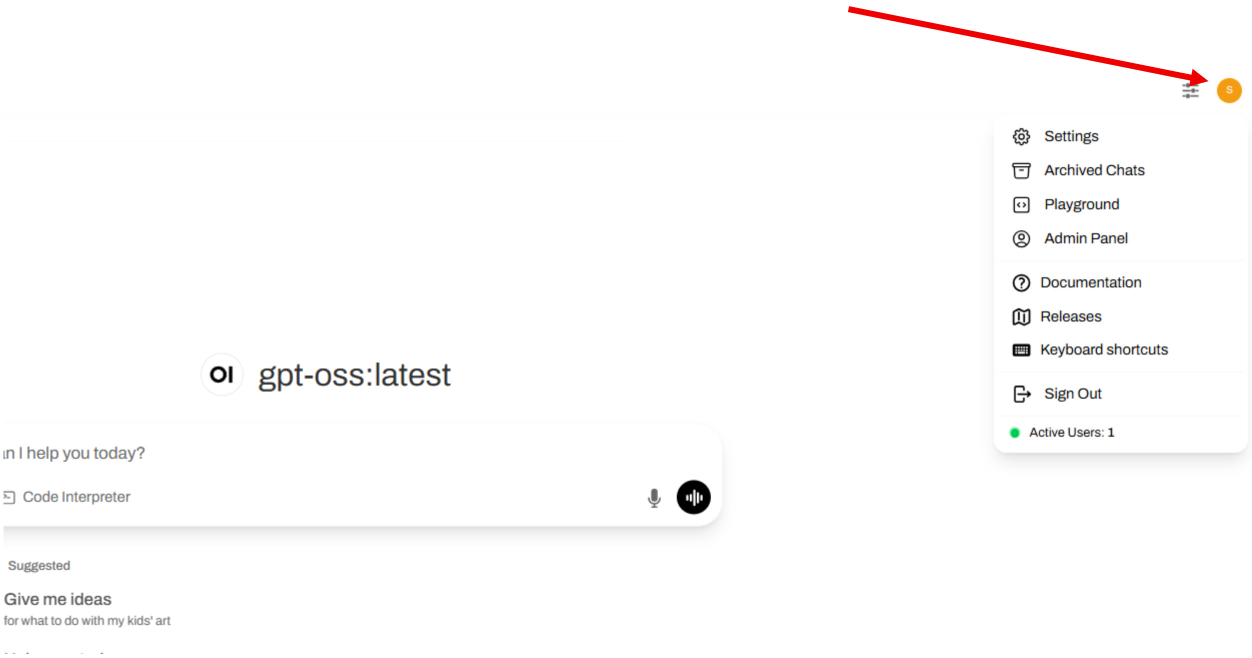
Settings you'll likely want (in Open WebUI). You will find all your configurations located in the settings tab.

Open **Settings** → **Advanced Parameters**:

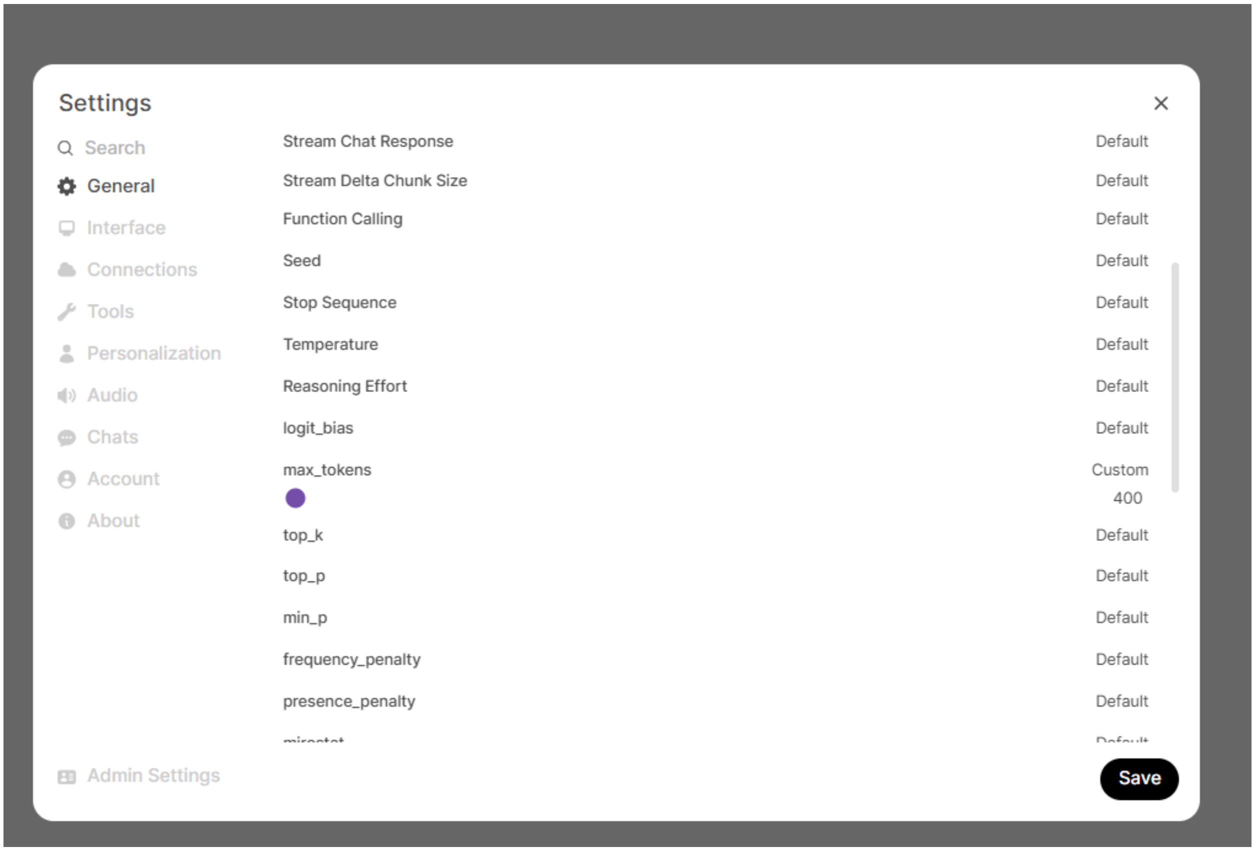
- **Max tokens** – controls maximum length of each reply.
  - If replies truncate, raise to **512–1024** (or more as needed).
- **Temperature / Top-p / Penalties** – creativity vs. accuracy.
- **Speed** – keep **Default** unless you specifically want faster/shorter outputs. Higher "speed" can reduce quality; prefer raising **Max tokens** instead.

After you create the admin account:

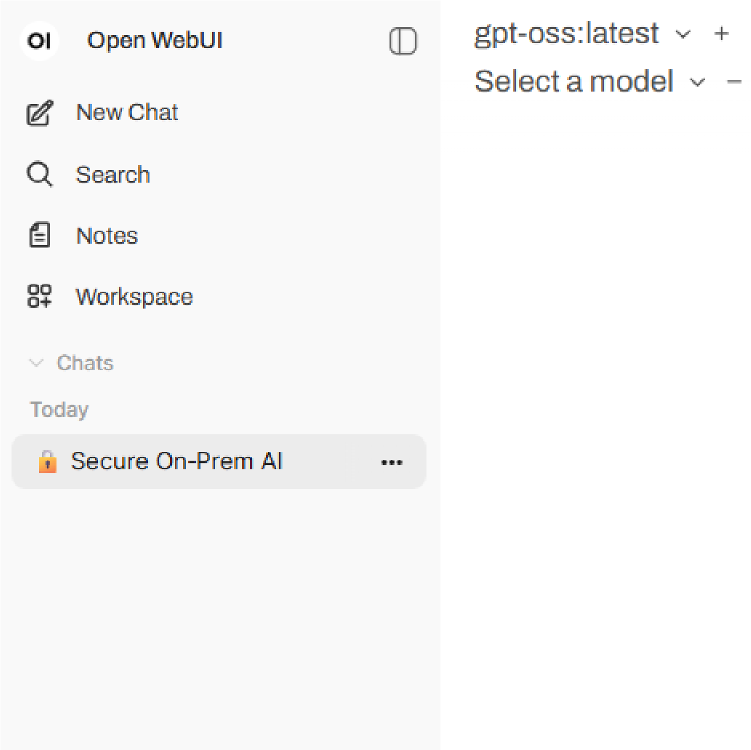
Open **Profile (top-right)** → **Admin Panel** → **Auth** and **disable new signups** if you want a private server.



Settings > Advanced Parameters



Stream Chat Response	Default
Stream Delta Chunk Size	Default
Function Calling	Default
Seed	Default
Stop Sequence	Default
Temperature	Default
Reasoning Effort	Default
logit_bias	Default
max_tokens	Custom 400
top_k	Default
top_p	Default
min_p	Default
frequency_penalty	Default
presence_penalty	Default
mirostat	Default



Other Notes:

Performance tips

- This AMI targets **g5.2xlarge (A10G)** class GPUs. Expect ~15–16 GB VRAM usage when the 20B model is hot.
- If responses feel slow under load:
  - Keep **one large model** active at a time.
  - Lower **Max tokens** to 256–512 for snappier replies.
  - Avoid running other GPU workloads on the instance.
- First response after a reboot or AMI launch is always slower (model warm-up).

Updating to newer images (optional)

```
cd /opt/agent-server

sudo docker compose pull

sudo docker compose up -d --force-recreate
```

OpenAI help: https://help.openai.com/en/collections/14446186-open-models-gpt-oss

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

AWS Data

- Data Encryption Configuration: This solution does not encrypt data within the running instance.
- User Credentials are stored: /root/.ssh/authorized_keys & /home/ubuntu/.ssh/authorized_keys
- Monitor the health:
    - Navigate to your Amazon EC2 console and verify that you're in the correct region.
    - Choose Instance and select your launched instance.
    - Select the server to display your metadata page and choose the Status checks tab at the bottom of the page to review if your status checks passed or failed.

Extra Information: (Optional)

Allocate Elastic IP

To ensure that your instance **keeps its IP during restarts** that might happen, configure an Elastic IP. From the EC2 console:

1. Select ELASTIC IPs.
2. Click on the ALLOCATE ELASTIC IP ADDRESS.
3. Select the default (Amazon pool of IPv4 addresses) and click on ALLOCATE.
4. From the ACTIONS pull down, select ASSOCIATE ELASTIC IP ADDRESS.
5. In the box that comes up, note down the Elastic IP Address, which will be needed when you configure your DNS.
6. In the search box under INSTANCE, click and find your INSTANCE ID and then click ASSOCIATE.
7. Your instance now has an elastic IP associated with it.
8. For additional help: https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/elastic-ip-addresses-eip.html