Private AI Workspace for DeepSeek and Local Models by Optick

Customer Instructions and Production Deployment Guide

Important first launch notice

Your first launch creates a unique administrator account, prepares the private workspace, downloads the included DeepSeek R1 7B model, and warms it on the NVIDIA GPU. **Allow up to 20 minutes.** The web page can appear before the workspace is ready. Do not reboot or try to sign in until the status command reports MODEL STATUS: READY.

1. What this product provides

This AWS Marketplace AMI deploys a private GPU-backed AI workspace in your AWS account. It includes Open WebUI for browser-based chat and Ollama for the local model runtime. The included DeepSeek R1 7B model runs locally on the instance GPU after the first-boot preparation completes.

- A customer-specific administrator account is created automatically at first boot.
- No external AI API key is required for the included DeepSeek R1 7B model.
- The Ollama service remains internal to the instance. TCP port 11434 is not published to the internet.
- The workspace is accessed from a browser over TCP port 80 using the instance public IPv4 address.
- First boot requires outbound internet access to complete the included model and workspace preparation.

2. Recommended AWS configuration

For the best customer experience, use a larger G5 instance rather than the smallest GPU profile. The product was validated on G5 GPU instances with NVIDIA A10G hardware.

Instance	GPU	CPU and memory	Recommendation
g5.xlarge	1 x NVIDIA A10G 24 GB	4 vCPU and 16 GiB	Smallest validated profile. Suitable for a cost-sensitive single-user test.
g5.4xlarge	1 x NVIDIA A10G 24 GB	16 vCPU and 64 GiB	Recommended production starting point. Faster first boot and more CPU and memory headroom.
g5.8xlarge	1 x NVIDIA A10G 24 GB	32 vCPU and 128 GiB	Use when the workspace also needs higher CPU or memory headroom. GPU capacity remains one A10G.

Storage recommendation: Start with 150 GiB gp3. Use 200 GiB gp3 or more when you plan to add models, store documents, or retain workspace data over time.

Important: Moving from g5.xlarge to g5.4xlarge improves CPU and memory headroom, but both have one A10G GPU with the same 24 GB GPU memory. Larger model capacity requires more GPU memory, not only more vCPU or RAM.

3. Launch and security group settings

1. Subscribe to the AWS Marketplace product and launch the AMI in a Region where G5 capacity is available.

2. Select g5.4xlarge for the recommended production experience. Select a key pair and enable a public IPv4 address when you need direct browser access.
3. Use at least 150 GiB gp3 storage. Increase storage before launch if you plan to add models or documents.
4. Configure the security group with these inbound rules:
    - TCP 22 from My IP only for temporary SSH administration.
    - TCP 80 from My IP during setup and testing. For a team deployment, restrict access to approved corporate IP ranges or a protected network path.
    - Do not open TCP 11434. The local Ollama service is intentionally not published to the host network.

For production team access, place the workspace behind a trusted network boundary such as a VPN, zero trust access proxy, or authenticated reverse proxy. Do not expose the browser interface broadly to the public internet.

4. First boot and readiness

The first boot is intentionally automated. It builds the initial workspace, creates your one-time administrator credentials, downloads the included model, and verifies GPU-backed model loading.

Be patient on the first boot. **Allow up to 20 minutes.** During this period, the browser page may load while the workspace is still preparing. Do not reboot the instance, remove containers, or attempt password resets while status is STARTING, DOWNLOADING, or WARMING.

Connect through SSH as **ubuntu** and run:

```
optick-status
```

Wait until the output shows:

```
MODEL STATUS: READY
```

The expected final signals are:

```
First-boot service: inactive
MODEL STATUS: READY
Local workspace health: PASS
```

To display the browser URL only, run:

```
optick-url
```

5. Sign in to the workspace

5. After MODEL STATUS: READY appears, retrieve the first-login details through SSH:

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

6. Open the displayed Workspace URL in a browser. The product works with the public IPv4 address and does not require a domain for initial access.
7. Sign in with the temporary administrator email and password from FIRST_LOGIN.txt.
8. Change the administrator password after your first successful sign-in.
9. Confirm that deepseek-r1:7b is selected and send a short prompt to verify your workspace.

Treat FIRST_LOGIN.txt as confidential. It contains the initial administrator password and is readable only by the **ubuntu** user on the instance.

6. Verify GPU-backed local inference

The included model is warmed during first boot. To confirm it is resident on the GPU, run:

```
sudo docker exec optick-ollama ollama ps
```

A healthy result shows deepseek-r1:7b with PROCESSOR set to 100% GPU. Ollama uses this command to show whether a loaded model is using GPU, CPU, or mixed memory.

To view host GPU information, run:

```
nvidia-smi
```

7. Everyday operations

Use the browser workspace for normal AI chat. The included local model is DeepSeek R1 7B. Additional compatible Ollama models may require more disk space and GPU memory; evaluate those requirements before adding models to a production workspace.

Useful status commands:

```
optick-status
optick-url
sudo docker exec optick-ollama ollama ps
nvidia-smi
```

Stop the EC2 instance when it is not needed to avoid ongoing EC2 compute charges. Your attached EBS storage remains available and continues to incur AWS storage charges while the instance is stopped.

8. Troubleshooting

If the first boot has not reached READY after 20 minutes, first confirm the instance is a supported G5 GPU instance and that outbound internet access is available. Then collect these read-only checks:

```
optick-status
sudo systemctl status optick-firstboot.service --no-pager
sudo journalctl -u optick-firstboot.service --no-pager -n 150
sudo docker exec optick-ollama ollama ps
nvidia-smi
```

Do not post the contents of FIRST_LOGIN.txt in a support request. Redact the administrator password before sharing any output.

9. Product resources

Official references for administration and security:

- Amazon EC2 G5 instance specifications
- Open WebUI Authentication and Access
- Open WebUI Hardening guidance
- Ollama FAQ and GPU verification
- AWS Marketplace AMI product guidance

10. Product support boundary

Seller support covers product launch guidance, first-boot readiness, workspace access, and basic product operation. AWS account setup, EC2 capacity availability, VPC design, corporate identity integration, custom model tuning, custom application development, and third-party model licensing remain the customer responsibility.