Headroom AI Context Compression Gateway by Code Creator

Customer Instructions for AWS Marketplace Users

Important first launch note: If the landing page or health check does not appear immediately after launch, wait one minute and refresh your browser. Headroom may take a short time to initialize on first startup.

1. What This AMI Provides

Headroom AI Context Compression Gateway by Code Creator provides a private, self hosted Headroom AI context compression gateway for AI agents, coding tools, LLM applications, OpenAI compatible clients, and Anthropic compatible clients. The AMI includes Headroom, Nginx, a public IP landing page, an examples page, first boot public IP refresh, health checks, and helper commands for day to day administration.

- Headroom AI context compression gateway installed as a system service.
- Public landing page and examples page available through Nginx on port 80.
- OpenAI compatible base URL and Anthropic compatible base URL examples.
- Telemetry disabled by default.
- No OpenAI, Anthropic, AWS, or Hugging Face API keys are stored in the AMI by default.
- Local Headroom service is bound to 127.0.0.1 and exposed through Nginx public routes.

2. Before You Begin

- Launch the instance from AWS Marketplace using this AMI.
- Use a security group that allows TCP 22 for SSH and TCP 80 for the web landing page.
- Have your Amazon EC2 private key available.
- Use the operating system user name ubuntu when connecting with SSH.
- Copy the public IPv4 address from the EC2 console after the instance is running.

Security note: This first launch configuration uses HTTP for simple public IP access. Do not send production provider API keys over plain HTTP unless you are using a trusted private network, VPN, SSH tunnel, or HTTPS configuration.

3. Connect to the Instance with SSH

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

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

Replace **PUBLIC_IP** with the public IPv4 address shown in the EC2 console.

4. First Login Instructions File

After logging in, view the generated customer instructions file:

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

This file is regenerated on first boot with the current instance public IP address. It includes the landing page, health check, examples page, client base URLs, and useful helper commands.

5. Browser Access

Open the landing page in your browser:

```
http://PUBLIC_IP/
```

Page or Endpoint	URL
Landing page	http://PUBLIC_IP/
Code Creator help page	http://PUBLIC_IP/codecreator.html
Client examples page	http://PUBLIC_IP/examples
Public Headroom health check	http://PUBLIC_IP/headroom-health
Public Headroom stats route	http://PUBLIC_IP/stats
OpenAI compatible base URL	http://PUBLIC_IP/v1
Anthropic compatible base URL	http://PUBLIC_IP

If the landing page does not appear immediately, wait one minute and refresh the browser. Also confirm that TCP 80 is allowed in the instance security group.

6. Useful Helper Commands

Command	Purpose
codecreator-headroom-url	Show public URLs and client base URLs.
codecreator-headroom-status	Show Headroom, Nginx, and local health status.
codecreator-headroom-health	Show full local Headroom health JSON.
codecreator-headroom-stats	Show Headroom stats, telemetry status, requests, tokens, and cost summary.
codecreator-headroom-examples	Show client connection examples from SSH.
codecreator-headroom-config	Interactively tune mode, budget, worker count, concurrency, and connection limits.
codecreator-headroom-logs	Show recent Headroom service logs.
codecreator-headroom-restart	Restart Headroom and verify health.

7. Configure OpenAI Compatible Clients

For OpenAI compatible tools and applications, use the base URL below. Replace **PUBLIC_IP** with your instance public IP address. Use your own provider API key in your client environment. The AMI does not store provider keys by default.

```
export OPENAI_BASE_URL=http://PUBLIC_IP/v1
export OPENAI_API_KEY=your_openai_api_key
```

Then configure your OpenAI compatible client, coding tool, or application to use the Headroom gateway as its base URL.

8. Configure Anthropic Compatible Clients and Claude Code

For Anthropic compatible tools such as Claude Code, use the public base URL below. Replace **PUBLIC_IP** with your instance public IP address.

```
export ANTHROPIC_BASE_URL=http://PUBLIC_IP
export ANTHROPIC_API_KEY=your_anthropic_api_key
```

Claude Code example:

```
ANTHROPIC_BASE_URL=http://PUBLIC_IP claude
```

9. Configure Headroom Runtime Settings

Use the configuration helper to tune the Headroom operating mode and safe runtime settings:

```
codecreator-headroom-config
```

- token mode prioritizes token compression and savings.
- cache mode prioritizes provider prefix cache stability.
- Daily budget is optional.
- API keys are not stored in the AMI by default.
- After saving changes, the helper restarts Headroom and checks health.

10. Health Checks and Service Verification

From SSH, check the local Headroom service:

```
curl http://127.0.0.1:8787/health | jq .
```

From a browser, use the public health route instead:

```
http://PUBLIC_IP/headroom-health
```

Why 127.0.0.1 does not work in your desktop browser: The address 127.0.0.1 means the local computer. It works from SSH inside the EC2 instance, but it does not point to the AWS server when typed into your desktop browser. Use http://PUBLIC_IP/headroom-health from your browser.

11. Restart or Inspect the Service

```
codecreator-headroom-status
codecreator-headroom-restart
codecreator-headroom-logs
```

The Headroom service is installed as a systemd service named **codecreator-headroom.service**. Nginx proxies public requests on port 80 to the local Headroom service.

12. Network Ports

Port	Purpose	Recommended Source
TCP 22	SSH access using the ubuntu user and your Amazon private key	Your trusted IP address, or 0.0.0.0/0 if required
TCP 80	Public landing page, examples	0.0.0.0/0 or your trusted network

	page, health route, and client base URL	
TCP 8787	Headroom local service only	Do not open publicly

13. Troubleshooting

Issue	Recommended Action
Landing page does not load	Wait one minute and refresh. Confirm the instance is running and TCP 80 is open in the security group.
Health check does not load in browser	Use http://PUBLIC_IP/headroom-health, not http://127.0.0.1:8787/health from your desktop browser.
SSH connection fails	Confirm TCP 22 is open to your IP, use the correct private key, and connect as the ubuntu user.
Headroom looks unhealthy	Run codecreator-headroom-status, then codecreator-headroom-logs, and restart with codecreator-headroom-restart.
Client cannot connect	Confirm the client is using http://PUBLIC_IP/v1 for OpenAI compatible clients or http://PUBLIC_IP for Anthropic compatible clients.
Concern about API keys over HTTP	Use a trusted private network, VPN, SSH tunnel, or configure HTTPS before sending production provider API keys.

14. Security and Production Notes

- Telemetry is disabled by default in this AMI.
- No provider API keys are stored in the AMI by default.
- The raw Headroom service is bound to 127.0.0.1:8787 and is not intended to be opened directly to the public internet.
- For production provider API usage, consider HTTPS, VPN, private networking, or SSH tunneling before sending sensitive keys or traffic.
- Keep the Ubuntu operating system updated according to your organization security policy.

15. Additional Resources

Headroom GitHub project: https://github.com/chopratejas/headroom

Headroom documentation: https://headroom-docs.vercel.app/docs/installation

Headroom PyPI package: https://pypi.org/project/headroom-ai/

AWS EC2 Linux instance connection guide:
https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/AccessingInstancesLinux.html