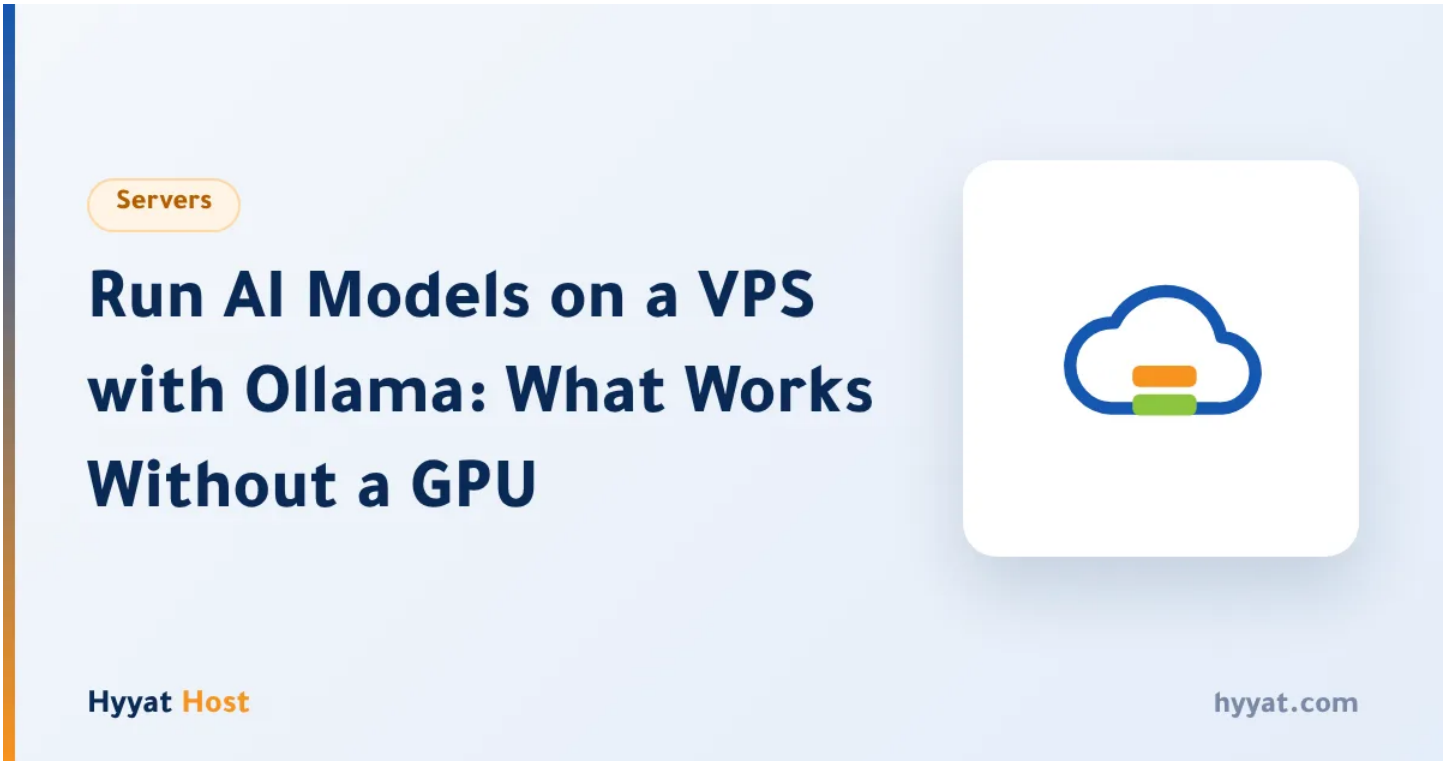
Run AI Models on a VPS with Ollama: What Works Without a GPU

Servers Without a Control Panel · https://www.hyyat.com/en/knowledgebase/run-ai-models-on-vps-ollama/

In short Ollama runs open language models on your own server, so your data stays with you. Our VPS servers have no **GPU**, so models run on the CPU: **small models** (up to about 3 billion parameters) are practical on an 8GB plan, and **mid-size models** need 16GB and are slower. Never expose the Ollama port to the internet.



Many businesses want an assistant that summarises documents or answers internal questions without sending data to an outside service. Ollama makes running open models a one-command job. It is important to know the limits first: these servers run on CPU only, so speed is much lower than large cloud services.

What to expect on CPU without a GPU

Model size	Approx. memory	Plan	Good for
Small (1 to 3 billion parameters)	2 to 4 GB	VPS 8GB	Summaries, classification, short replies
Medium (7 to 8 billion parameters)	5 to 8 GB	VPS 16GB	Better answers, noticeably slower
Large (over 13 billion)	Over 10 GB	Not practical without a GPU	Use a cloud service or a GPU server

Figures are approximate for the quantised models Ollama uses by default, and vary with the model and text length.

1. Install Ollama

```
curl -fsSL https://ollama.com/install.sh | sh
systemctl status ollama
```

2. Try a small model

```
ollama pull qwen2.5:3b
ollama run qwen2.5:3b "Summarise this paragraph in two lines: ..."
```

The Qwen family handles Arabic well compared with other small models. Test several models on your real texts before choosing.

3. A team interface (optional)

Open WebUI is a chat-style interface that works with Ollama. Run it with Docker and keep it behind HTTPS with a login:

```
docker run -d --name open-webui --restart unless-stopped \
  --add-host=host.docker.internal:host-gateway \
```

```
  -e OLLAMA_BASE_URL=http://host.docker.internal:11434 \
  -v open-webui:/app/backend/data -p 127.0.0.1:3000:8080 \
  ghcr.io/open-webui/open-webui:main
```

Then point a subdomain at port 3000 through Caddy, as in the n8n guide, and create the admin account straight away.

A mistake to avoid: Opening the Ollama port (11434) to the internet. It has no password, so anyone could use your server and its resources. Keep it on 127.0.0.1 and give access through a protected interface.

At Hyyat Host: VPS servers with up to 16GB of memory to test open models on your own data: VPS plans, with our own control panel for power, console and snapshots, and Arabic and English support.

Frequently asked questions

Do I need a GPU to run AI models?

Not for small models, which run on the CPU at an acceptable speed. For large models or heavy use, a GPU becomes necessary.

Does my data leave the server?

No. Ollama runs the model on your server, and your text is not sent to any outside service.

Is it suitable for high-volume customer chat?

On CPU only, it suits internal use or a limited number of concurrent requests. Heavy load needs more resources or a specialised service.

Related guides

- New VPS Setup Commands, Step by Step
- Docker Compose Generator for WordPress and Apps
- Firewall: Control the Open Ports on Your VPS
- When Should You Move from Shared Hosting to a VPS? 7 Clear Signs

All guides for servers without a panel →

Need this service from Hyyat Host?Our team handles it for you, with support replying to tickets within 24 hours at most.

Server Management