

????????????????AI

????????????????

???????? 30???? Linux Docker????
???????? AI????

????????????????

????????????

1. ?????????????

??

80%??

????????

| ?? | ?? | ???? |
|--------|------|------|
| ?? | ★ | √ |
| ?? | ★ | √ |
| PPT?? | ★★ | √ |
| ???? | ★★ | √ |
| ???? | ★★ | √ |
| ???? | ★ | √ |
| ???? | ★★ | √ |
| ???? | ★★ | √ |
| SQL?? | ★★ | √ |
| ???? | ★★★ | √ |
| OA???? | ★★★★ | ?? |
| ERP?? | ★★★★ | ?? |
| CRM?? | ★★★★ | ?? |

□□□

????

□□□□□□□□

□□□

- □□□□ Agent
- Dify
- FastGPT
- Coze
- MaxKB

□□□

??????

□□□□□□

100???

□□□

??A

□□□□ + □□□□

□□□



└─ PPT Agent

└─ Agent

□□□

- □□□ GPU
- □□□□□
- □□□□

1□□□□□

????????????AI??

□□□□

??B

Dify + □□□□ API

□□□

```
□□
|
▼
Dify
|
└─ Qwen3
└─ DeepSeek
└─ Claude
└─ GPT
|
└─ □□□□□
```

□□□

1

□□ Agent□□□□

□□□

AI□□□□

AI□□□□

AI□□□□

AI□□□□

AI□□□□

□□□□□□

2

□□ RAG□□□

□□□

Word

PDF

Excel

□□□□

□□□□

□□□□

SOP□□

□□□

3

□□□□

□□□□

□□□□□□

??????Agent????

□□□□□□□□

□□□□

□□□□.pdf

Agent□□□□

□□□□
□□□□
□□□
□□□□
□□□□
□□□□
□□□□

□□□

□□□
□□□
□□□

□□□□

Dify
+
Qwen-Max

+

□□□□

□□

□□□□□

???????Agent

??

□□

- □□□
- Flux
- Midjourney

□□□□

□□□□

□□□

□□□

□□□□

□□□□

PPT□□

??

□□

- □□□□
- □□
- Kling

□□□

□□□□

□□□□

□□□□

????????????

□□□□

?????

□□□□ GPU

□□□□ API□

□□□

□□□ Qwen3 235B□

□□□□

□□□

4 × A800

□

4 × H20

□□□

50□~100□+

□□□□

□□

□□

□□

□□

□□□□

?????

□□□□□□

□□□

200+

□□

□□□

5000+

□□

□□□□□□

??

DeepSeek-R1 70B

Qwen3 32B

Qwen3 72B

□□□

Ubuntu 24.04

Docker

vLLM

Ollama

OpenWebUI

????????????

□□□□□□□□□□

□□1□1□□□

□□□□□

+

Dify

+

□□□□□

□□□

□□□□

□□□□

□□□□

PPT□□

□□□□

□□2□3□□□

□□SS0□□

□□□□□

□□□□

□□□□

□□3□□□□

□□□□□□□

OpenWebUI

+

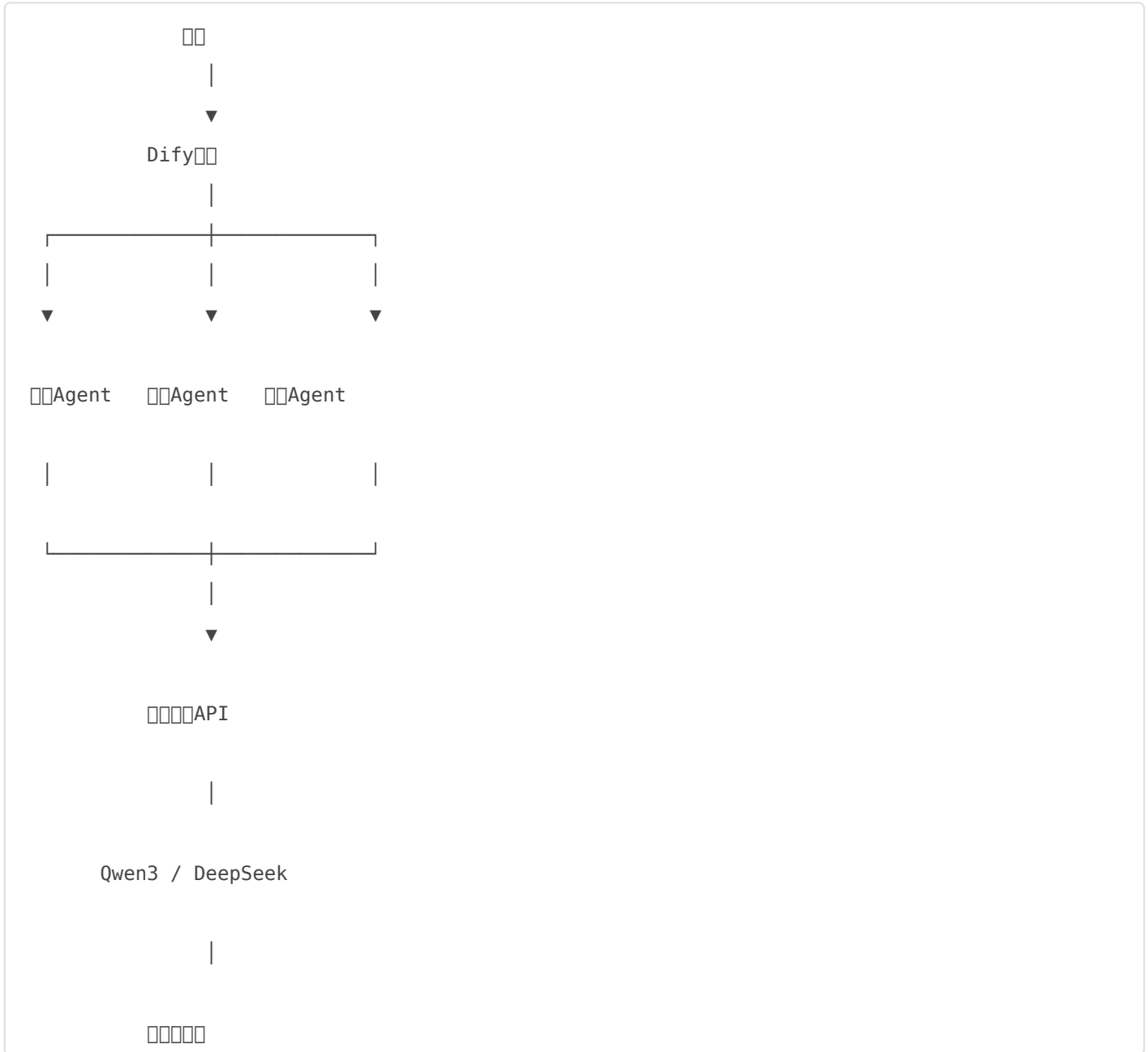
Dify

+

Qwen3

+

???????



Dify OpenWebUI

[Milvus](#) [Qdrant](#)



████████████████

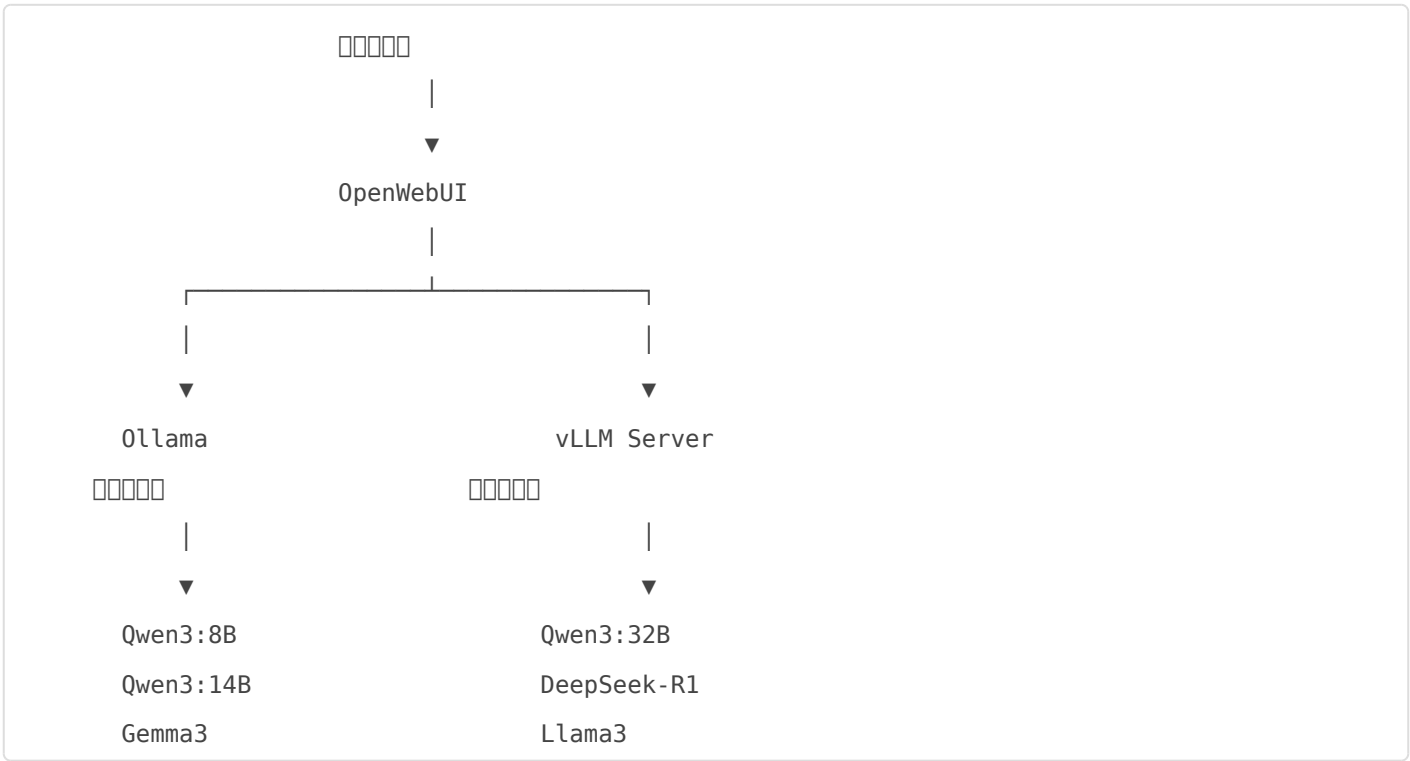
AI██████

Ubuntu 24.04
 Docker
 vLLM
 Ollama
 OpenWebUI

██

???????

████████



███

| ██ | ██ |
|-----------|-------------|
| OpenWebUI | ChatGPT████ |
| Ollama | ████ |
| vLLM | ████ |
| Docker | ██ |
| Nginx | ███ |

| □□ | □□ |
|------------|-----|
| PostgreSQL | □□□ |
| Redis | □□ |

??????

????

□□□□

CPU□
32 Core

□□□
128GB

GPU□
RTX4090 24GB

□□□□
1TB NVMe

□□□□

Qwen3:8B
Qwen3:14B
DeepSeek-R1:8B

????

□□□□

CPU□
64 Core

□□□

256GB

GPU□

2 × RTX4090

□□□□

Qwen3 32B

DeepSeek-R1 32B

???? Docker

□□□□

```
apt update
```

```
apt upgrade -y
```

□□□□□

```
apt install -y \
```

```
curl \
```

```
wget \
```

```
git \
```

```
vim \
```

```
unzip
```

□□ Docker

```
curl -fsSL https://get.docker.com | bash
```

□□

```
docker version
```

□□ docker□

```
usermod -aG docker $USER
```

□□□□

???? NVIDIA ??

□□□□

```
lspci | grep NVIDIA
```

□□□□

```
ubuntu-drivers autoinstall
```

□□

```
reboot
```

□□

```
nvidia-smi
```

□□□□□

```
Driver Version: 575.xx  
CUDA Version: 12.x
```

???? NVIDIA Container Toolkit

□□□□

```
curl -fsSL https://nvidia.github.io/libnvidia-container/gpgkey \  
| gpg --dearmor \  
-o /usr/share/keyrings/nvidia-container-toolkit-keyring.gpg
```

□□

```
apt install -y nvidia-container-toolkit
```

□□

```
nvidia-ctl runtime configure --runtime=docker
```

☐ docker

```
systemctl restart docker
```

☐

```
docker run --rm \  
--gpus all \  
nvidia/cuda:12.4.1-runtime-ubuntu22.04 \  
nvidia-smi
```

???? Ollama

☐☐☐

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

☐

```
systemctl status ollama
```

☐☐☐

☐☐

```
/etc/systemd/system/ollama.service
```

☐☐

```
Environment="OLLAMA_HOST=0.0.0.0:11434"
```

☐

```
systemctl daemon-reload
```

```
systemctl restart ollama
```

☐

```
ss -lntp | grep 11434
```

??????

□□

```
ollama pull qwen3:8b
```

□

```
ollama pull deepseek-r1:8b
```

□□

```
ollama list
```

□□

```
ollama run qwen3:8b
```

???? vLLM

□□ Docker□□

□□□□

```
mkdir -p /data/vllm
```

□□

```
docker run -d \  
  --name vllm \  
  --gpus all \  
  -p 8000:8000 \  
  -v /data/models:/models \  
  vllm/vllm-openai:latest \  
  --model /models/Qwen3-32B
```


????? vLLM

OpenWebUI

Admin Settings

↓

Connections

↓

OpenAI API

URL:

http://IP:8000/v1

API KEY:

EMPTY

????? HTTPS

Ngix

```
apt install nginx -y
```

```
server {
```

```
    listen 443 ssl;
```

```
    server_name ai.company.com;
```

```
    location / {
```

```
        proxy_pass http://127.0.0.1:3000;
```

```
    proxy_set_header Host $host;
    proxy_set_header X-Real-IP $remote_addr;

}

}
```

□□

```
nginx -t

systemctl reload nginx
```

????????????

□□□□□□□□□□

Docker Compose □□

□□□□

```
/data/ai-platform
```

```
├─ openwebui
├─ ollama
├─ vllm
├─ postgres
├─ redis
├─ nginx
└─ backup
```

□□□□

```
PostgreSQL
Redis
MinIO
OpenWebUI
Ollama
vLLM
Nginx
```

