RagFlow

notes:

add to /etc/sysctl.conf:

```
vm.max_map_count=262144
sudo sysctl -w vm.max_map_count=262144
```

```
git clone https://github.com/infiniflow/ragflow.git

cd ragflow/docker
chmod +x ./entrypoint.sh
```

adjust docker-compose.yml, add

```
    deploy:
      resources:
        reservations:
          devices:
          - driver: nvidia
            device_ids: ['0']
            capabilities: [gpu]
```

for change ports

```
docker compose up -d
```

Docker ENV Variables...: user_default_llm

Newly signed-up users use LLM configured by this part. Otherwise, user need to configure his own LLM in setting. factory

The LLM suppliers. “OpenAI”， “Tongyi-Qianwen”, “ZHIPU-AI”, “Moonshot”, “DeepSeek”, “Baichuan”, and “VolcEngine” are supported. api_key The corresponding API key of your assigned LLM vendor.

ollama docker compose:

```
docker compose -f docker-ollama.yml up -d
```

run models (more on https://ollama.com/library): preferred ones: mistral-openorca , llama3-chatqa?

```
llama3, qwen2, phi3, mistral, mixtral, codegemma, codellama, dolphin-
mixtral, deepseek-coder, mistral-openorca , codestral, codeqwen, orca2,
llama3-gradient,  llama3-chatqa (nvidia for RAG) and others...
```

download/update model:

```
docker exec -it ollama ollama pull codestral
```

list available models:

```
docker exec -it ollama ollama list
```

run model locally:

```
docker exec -it ollama ollama run llama3
```