

# Linux + IA 2025/05/27

## GGUF

- Crear carpeta gguf

- `mkdir gguf && cd gguf`

- Descargar:

- `wget -c https://nextcloud.openit.dev/s/4dRA4KdJb7rqDqw/download/Llama-3.2-3B-Instruct-Q4_K_M.gguf`

## LLAMA.CPP

- SITO <https://github.com/ggml-org/llama.cpp>
- Se debe estar en la carpeta %HOME

- `cd`

- Descarga

- `wget -c https://github.com/ggml-org/llama.cpp/releases/download/b9351/llama-b9351-bin-ubuntu-x64.tar.gz`

- `wget -c https://github.com/ggml-org/llama.cpp/releases/download/b9371/llama-b9371-bin-ubuntu-x64.tar.gz`  
`tar -xvzf llama-b9371-bin-ubuntu-x64.tar.gz`

- Crear carpeta build/ y dentro pasar la carpeta llama-b9351 renombrandola bin/

- `mkdir build && mv llama-b9371/ build/bin/`

- Agregar PATH

- `nano .bashrc`

- Agregar al final

- `export PATH="$HOME/build/bin:$PATH"`

- Aplicar los parametros en la sesión

- `source .bashrc`

- Verificación

- `llama-cli --version`

- Ejecutandolo como servidor

- `llama-server -m gguf/Llama-3.2-3B-Instruct-Q4_K_M.gguf --host 0.0.0.0 --port 11435 -c 2048 -t 4`

- Script de inicialización

- [start-llama.sh](#)

```
#!/bin/bash

# Cambiar esta ruta cuando quieras otro modelo
MODEL="$HOME/gguf/Llama-3.2-3B-Instruct-Q4_K_M.gguf"

nohup llama-server \
  --jinja \
  -m "$MODEL" \
  --host 127.0.0.1 \
  --port 11435 \
  -c 8192 \
  -t 4 \
  > /tmp/llama-server.log 2>&1 &

echo $! > /tmp/llama-server.pid
echo "Servidor iniciado. PID: $(cat /tmp/llama-server.pid)"
```

- Permisos y ejecutar

- `chmod u+x start-llama.sh`  
`./start-llama.sh`

- Script finalización

- [stop-llama.sh](#)

```
#!/bin/bash

if [ -f /tmp/llama-server.pid ]; then
  kill $(cat /tmp/llama-server.pid)
  rm /tmp/llama-server.pid
  echo "Servidor detenido"
else
  pkill -f llama-server
  echo "Servidor detenido"
fi
```

- Permisos y ejecutar

- `chmod u+x stop-llama.sh`  
`./stop-llama.sh`

## AGENTE

- Instalación

- `pipx install git+https://github.com/OpenInterpreter/open-interpreter.git`

- Uso con ollama

- `interpreter --local`

- Uso con llama.cpp

- `interpreter --api_base "http://127.0.0.1:11435/v1" --api_key "dummy"`

From:

<https://dw.openit.dev/> - **DokuWikiOpenIT**

Permanent link:

<https://dw.openit.dev/lia20260527>

Last update: **2026/05/28 01:15**

