

2026/08/24

IA Local

- Sitio donde esta los ultimos compilados: <https://github.com/ggml-org/llama.cpp/releases>
- Descargarlo:

wget

```
https://github.com/ggml-org/llama.cpp/releases/download/b4610/llama-b4610-bin-ubuntu-x64.zip
```

- Otro tar.gz

- `wget -c https://github.com/ggml-org/llama.cpp/releases/download/b10615/llama-b10615-bin-ubuntu-x64.tar.gz`

- Descomprimir:

- `unzip llama-b4610-bin-ubuntu-x64.zip`

- Descargando GGUF Gemma

- `wget -O Gemma-2-2B-IT-Platinum-Q4_K_M.gguf \`
`https://huggingface.co/CelesteImperia/Gemma-2-2B-IT-GGUF/resolve/main/Gemma-2-2B-IT-Platinum-Q4_K_M.gguf?download=true`

- qwen

- `wget -O qwen2.5-0.5b-instruct-q4_k_m.gguf \`
`"https://huggingface.co/Qwen/Qwen2.5-0.5B-Instruct-GGUF/resolve/main/qwen2.5-0.5b-instruct-q4_k_m.gguf"`

Ejecutado llama.cpp

- llama-cli, OJO tiene que estar dentro de la carpeta precompilada.

- `./llama-cli -m ../qwen2.5-0.5b-instruct-q4_k_m.gguf --ctx-size 2048 --temp 0.7 --repeat-penalty 1.1 -n -1 -p "que modelo eres"`

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

Permanent link:
<https://dw.openit.dev/lsf-20260824>

Last update: 2026/08/25 01:23



