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`