Ampliación de Fundamentos de Hardware

Angel Fabrizio Ullaguari Yanza

2ºASIR

Práctica IA. Crear LLM

Objetivo

Crear un LLM especializado llamado "Chef-bot". Su única función será recibir una lista de ingredientes y generar una receta sencilla en español. No será un chatbot general; será un experto en cocina.

Requisitos

- Modelo Base: microsoft/Phi-3-mini-4k-instruct (potente, pequeño y con una licencia permisiva).
- Dataset: Crearemos un pequeño dataset de recetas para enseñarle su nueva habilidad.
- Herramientas: Usaremos Axolotl para el fine-tuning y llama.cpp para la cuantización final.

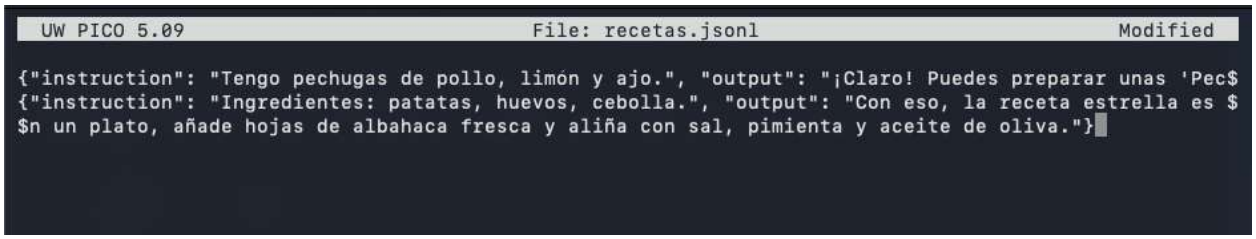
Configuración

Creación del directorio mi-chef-bot y también creamos el entorno virtual con python.

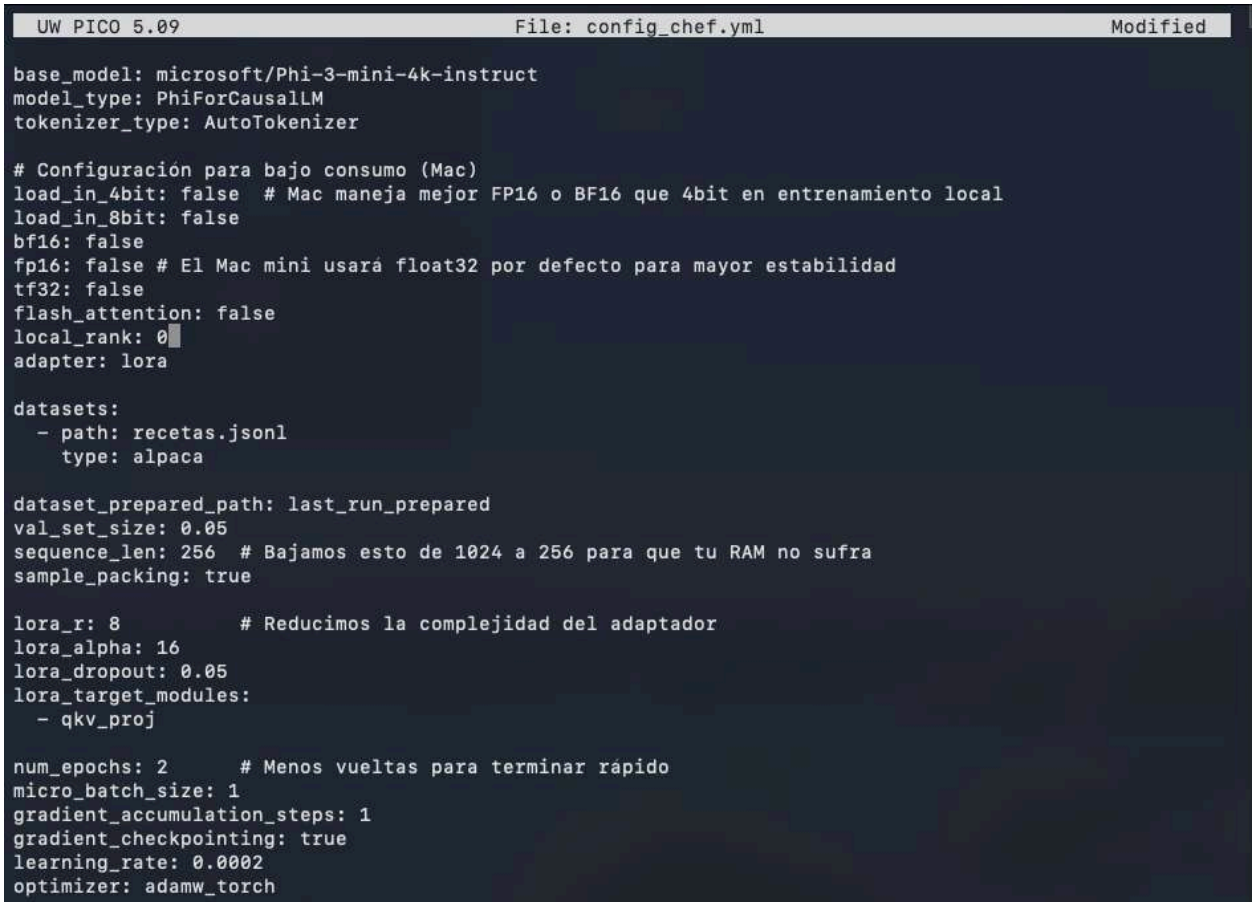
```
(venv) aulaateca26@Mac-mini-de-aulaateca26 mi-chef-bot % rm -rf venv
(venv) aulaateca26@Mac-mini-de-aulaateca26 mi-chef-bot % python3.11 -m venv venv
(venv) aulaateca26@Mac-mini-de-aulaateca26 mi-chef-bot % source venv/bin/activate
(venv) aulaateca26@Mac-mini-de-aulaateca26 mi-chef-bot % pip install "axolotl @ git+https://github.com
/OpenAccess-AI-Collective/axolotl"
Collecting axolotl @ git+https://github.com/OpenAccess-AI-Collective/axolotl
  Cloning https://github.com/OpenAccess-AI-Collective/axolotl to /private/var/folders/x7/5vw0c3rx46j0b
hmp88zdy3zm0000gn/T/pip-install-0o_ika1v/axolotl_c6c119a320f94583bc3134b9ecc3fdd0
  Running command git clone --filter=blob:none --quiet https://github.com/OpenAccess-AI-Collective/axo
lotl /private/var/folders/x7/5vw0c3rx46j0bhmp88zdy3zm0000gn/T/pip-install-0o_ika1v/axolotl_c6c119a320f
94583bc3134b9ecc3fdd0
  Resolved https://github.com/OpenAccess-AI-Collective/axolotl to commit 236dad3bb7954e5dc8dcca5b4d067
a7e6e39fb7e
  Installing build dependencies ... done
  Getting requirements to build wheel ... done
  Preparing metadata (pyproject.toml) ... done
Collecting packaging==26.0 (from axolotl @ git+https://github.com/OpenAccess-AI-Collective/axolotl)
  Using cached packaging-26.0-py3-none-any.whl.metadata (3.3 kB)
Collecting huggingface_hub>=1.1.7 (from axolotl @ git+https://github.com/OpenAccess-AI-Collective/axol
otl)
  Using cached huggingface_hub-1.4.0-py3-none-any.whl.metadata (13 kB)
Collecting peft>=0.18.1 (from axolotl @ git+https://github.com/OpenAccess-AI-Collective/axolotl)
  Using cached peft-0.18.1-py3-none-any.whl.metadata (14 kB)
Collecting tokenizers>=0.22.1 (from axolotl @ git+https://github.com/OpenAccess-AI-Collective/axolotl)
  Using cached tokenizers-0.22.2-cp39-abi3-macosx_11_0_arm64.whl.metadata (7.3 kB)
Collecting transformers==5.0.0 (from axolotl @ git+https://github.com/OpenAccess-AI-Collective/axolotl
)
  Using cached transformers-5.0.0-py3-none-any.whl.metadata (37 kB)
Collecting accelerate==1.12.0 (from axolotl @ git+https://github.com/OpenAccess-AI-Collective/axolotl)
  Using cached accelerate-1.12.0-py3-none-any.whl.metadata (19 kB)
Collecting datasets==4.5.0 (from axolotl @ git+https://github.com/OpenAccess-AI-Collective/axolotl)
  Using cached datasets-4.5.0-py3-none-any.whl.metadata (19 kB)
Collecting trl==0.27.1 (from axolotl @ git+https://github.com/OpenAccess-AI-Collective/axolotl)
  Using cached trl-0.27.1-py3-none-any.whl.metadata (11 kB)
Collecting hf_xet==1.2.0 (from axolotl @ git+https://github.com/OpenAccess-AI-Collective/axolotl)
  Using cached hf_xet-1.2.0-cp37-abi3-macosx_11_0_arm64.whl.metadata (4.9 kB)
Collecting kernels==0.11.5 (from axolotl @ git+https://github.com/OpenAccess-AI-Collective/axolotl)
  Using cached kernels-0.11.5-py3-none-any.whl.metadata (3.3 kB)
Collecting trackio>=0.13.0 (from axolotl @ git+https://github.com/OpenAccess-AI-Collective/axolotl)
  Using cached trackio-0.15.0-py3-none-any.whl.metadata (9.9 kB)
Collecting typing-extensions>=4.15.0 (from axolotl @ git+https://github.com/OpenAccess-AI-Collective/a
xolotl)
  Using cached typing_extensions-4.15.0-py3-none-any.whl.metadata (3.3 kB)
Collecting optimum==1.16.2 (from axolotl @ git+https://github.com/OpenAccess-AI-Collective/axolotl)
  Using cached optimum-1.16.2-py3-none-any.whl.metadata (17 kB)
Collecting hf_transfer (from axolotl @ git+https://github.com/OpenAccess-AI-Collective/axolotl)
  Using cached hf_transfer-0.1.9-cp38-abi3-macosx_11_0_arm64.whl.metadata (1.7 kB)
Collecting sentencepiece (from axolotl @ git+https://github.com/OpenAccess-AI-Collective/axolotl)
  Downloading sentencepiece-0.2.1-cp311-cp311-macosx_11_0_arm64.whl.metadata (10 kB)
Collecting gradio<7.0,>=6.2.0 (from axolotl @ git+https://github.com/OpenAccess-AI-Collective/axolotl)
```

.

Ahora creamos el archivo recetas.json con nano, y ponemos los prompts pedidos en el enunciado.



También crearemos un archivo .yml para configurar el modelo phi-3-mini (modelo chef-bot).



Ahora clonamos llama.cpp con Github.

```
(venv) aulaateca26@Mac-mini-de-aulaateca26 mi-chef-bot % ls
config_chef.yml recetas.jsonl   venv
(venv) aulaateca26@Mac-mini-de-aulaateca26 mi-chef-bot % git clone https://github.com/ggerganov/llama.cpp
Clonando en 'llama.cpp'...
remote: Enumerating objects: 78196, done.
remote: Counting objects: 100% (89/89), done.
remote: Compressing objects: 100% (65/65), done.
remote: Total 78196 (delta 53), reused 24 (delta 24), pack-reused 78107 (from 3)
Recibiendo objetos: 100% (78196/78196), 285.39 MiB | 44.84 MiB/s, listo.
Resolviendo deltas: 100% (56583/56583), listo.
(venv) aulaateca26@Mac-mini-de-aulaateca26 mi-chef-bot % cd llama.cpp
(venv) aulaateca26@Mac-mini-de-aulaateca26 llama.cpp % make
Makefile:6: *** Build system changed:
 The Makefile build has been replaced by CMake.

 For build instructions see:
 https://github.com/ggml-org/llama.cpp/blob/master/docs/build.md

.  Stop.
(venv) aulaateca26@Mac-mini-de-aulaateca26 llama.cpp %
```

E instalamos las librerias en llama.cpp

```
(venv) aulaateca26@Mac-mini-de-aulaateca26 llama.cpp % pip install -r requirements.txt
Looking in indexes: https://pypi.org/simple, https://download.pytorch.org/whl/cpu, https://download.pytorch.org/whl/nightly, https://download.pytorch.org/whl/cpu, https://download.pytorch.org/whl/nightly, https://download.pytorch.org/whl/cpu, https://download.pytorch.org/whl/nightly
Ignoring torch: markers 'platform_machine == "s390x"' don't match your environment
Ignoring torch: markers 'platform_machine == "s390x"' don't match your environment
Collecting numpy~=1.26.4 (from -r /Users/aulaateca26/mi-chef-bot/llama.cpp/requirements/requirements-convert_legacy_llama.txt (line 1))
  Using cached numpy-1.26.4-cp311-cp311-macosx_11_0_arm64.whl.metadata (114 kB)
Requirement already satisfied: sentencepiece~=0.2.0 in /Users/aulaateca26/mi-chef-bot/venv/lib/python3.11/site-packages (from -r /Users/aulaateca26/mi-chef-bot/llama.cpp/requirements/requirements-convert_legacy_llama.txt (line 2)) (0.2.1)
Collecting transformers<5.0.0,>=4.57.1 (from -r /Users/aulaateca26/mi-chef-bot/llama.cpp/requirements/requirements-convert_legacy_llama.txt (line 4))
  Using cached transformers-4.57.6-py3-none-any.whl.metadata (43 kB)
Collecting gguf>=0.1.0 (from -r /Users/aulaateca26/mi-chef-bot/llama.cpp/requirements/requirements-convert_legacy_llama.txt (line 6))
  Using cached https://download.pytorch.org/whl/nightly/gguf-0.17.1-py3-none-any.whl.metadata (4.3 kB)
Collecting protobuf<5.0.0,>=4.21.0 (from -r /Users/aulaateca26/mi-chef-bot/llama.cpp/requirements/requirements-convert_legacy_llama.txt (line 7))
  Using cached protobuf-4.25.8-cp37-abi3-macosx_10_9_universal2.whl.metadata (541 bytes)
Collecting torch~=2.6.0 (from -r /Users/aulaateca26/mi-chef-bot/llama.cpp/requirements/requirements-convert_hf_to_gguf.txt (line 5))
  Using cached https://download.pytorch.org/whl/cpu/torch-2.6.0-cp311-none-macosx_11_0_arm64.whl.metadata (28 kB)
Collecting aiohttp~=3.9.3 (from -r /Users/aulaateca26/mi-chef-bot/llama.cpp/requirements/requirements-tool_bench.txt (line 1))
  Using cached https://download.pytorch.org/whl/nightly/aiohttp-3.9.5-cp311-cp311-macosx_11_0_arm64.whl (390 kB)
```

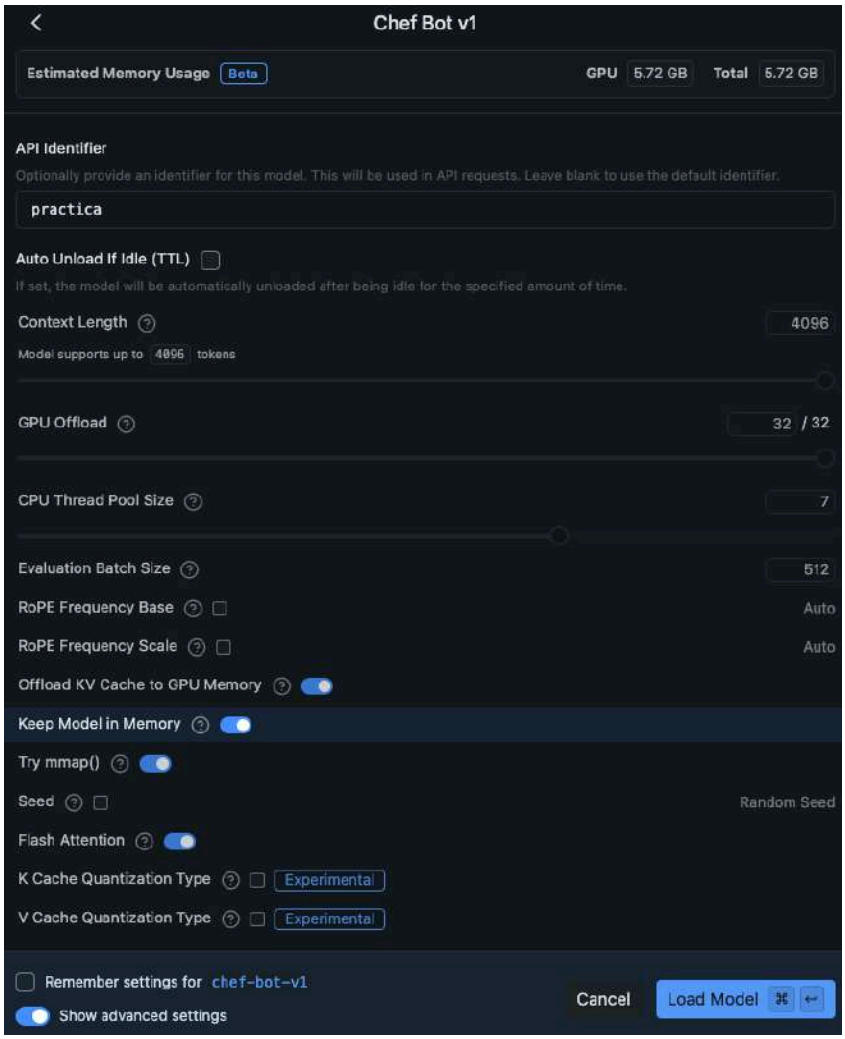
Ahora, instalamos git lfs en llama.cpp, que sirve para gestionar modelos grandes de tamaño.

```
(venv) aulaateca26@Mac-mini-de-aulaateca26 llama.cpp % git lfs install
Updated Git hooks.
Git LFS initialized.
(venv) aulaateca26@Mac-mini-de-aulaateca26 llama.cpp % git clone https://huggingface.co/microsoft/Phi-3-mini-4k-instruct
Clonando en 'Phi-3-mini-4k-instruct'...
remote: Enumerating objects: 111, done.
remote: Counting objects: 100% (6/6), done.
remote: Compressing objects: 100% (6/6), done.
remote: Total 111 (delta 1), reused 0 (delta 0), pack-reused 105 (from 1)
Recibiendo objetos: 100% (111/111), 1.03 MiB | 3.99 MiB/s, listo.
Resolviendo deltas: 100% (52/52), listo.
Filtrando contenido: 100% (3/3), 3.11 GiB | 24.01 MiB/s, listo.
(venv) aulaateca26@Mac-mini-de-aulaateca26 llama.cpp %
```

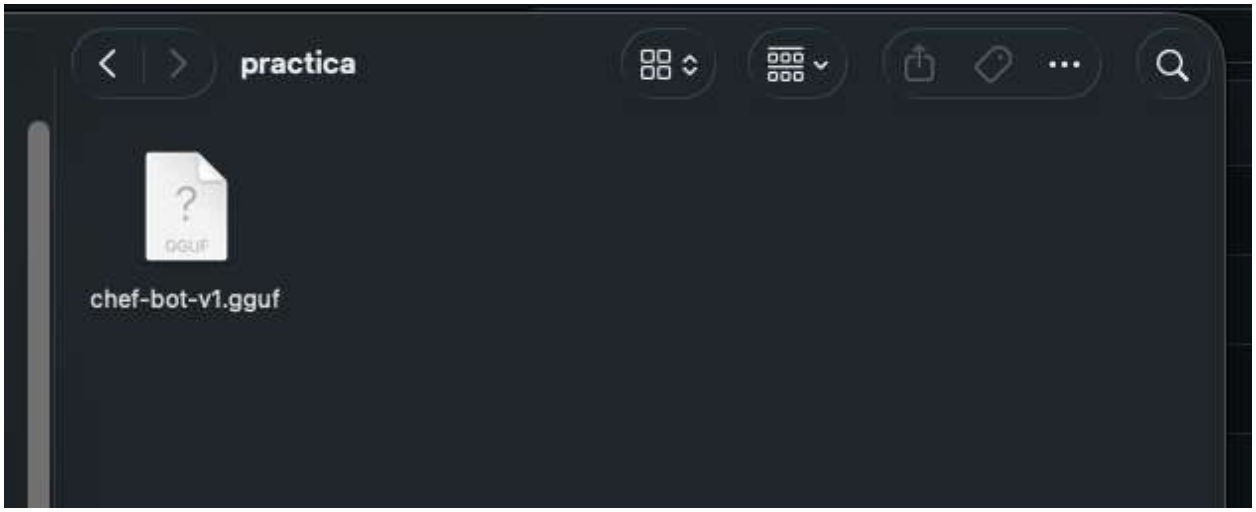
Y luego, convertimos el modelo en gguf para que pueda utilizarse en LM Studio.

```
(venv) aulaateca26@Mac-mini-de-aulaateca26 llama.cpp % ls -d ../*/
../llama.cpp/                  ../Phi-3-mini-4k-instruct/      ../venv/
(venv) aulaateca26@Mac-mini-de-aulaateca26 llama.cpp % python convert_hf_to_gguf.py ../Phi-3-mini-4k-instruct --outfile ../chef-bot-v1.gguf --outtype q8_0
INFO:hf-to-gguf:Loading model: Phi-3-mini-4k-instruct
INFO:hf-to-gguf:Model architecture: Phi3ForCausalLM
INFO:hf-to-gguf:gguf: loading model weight map from 'model.safetensors.index.json'
INFO:hf-to-gguf:gguf: indexing model part 'model-00001-of-00002.safetensors'
INFO:hf-to-gguf:gguf: indexing model part 'model-00002-of-00002.safetensors'
INFO:gguf.gguf_writer:gguf: This GGUF file is for Little Endian only
INFO:hf-to-gguf:Exporting model...
INFO:hf-to-gguf:token_embd.weight,         torch.bfloat16 --> Q8_0, shape = {3072, 32064}
INFO:hf-to-gguf:blk.0.attn_norm.weight,    torch.bfloat16 --> F32, shape = {3072}
INFO:hf-to-gguf:blk.0.ffn_down.weight,     torch.bfloat16 --> Q8_0, shape = {8192, 3072}
INFO:hf-to-gguf:blk.0.ffn_up.weight,       torch.bfloat16 --> Q8_0, shape = {3072, 16384}
INFO:hf-to-gguf:blk.0.ffn_norm.weight,     torch.bfloat16 --> F32, shape = {3072}
INFO:hf-to-gguf:blk.0.attn_output.weight,  torch.bfloat16 --> Q8_0, shape = {3072, 3072}
INFO:hf-to-gguf:blk.0.attn_qkv.weight,     torch.bfloat16 --> Q8_0, shape = {3072, 9216}
INFO:hf-to-gguf:blk.1.attn_norm.weight,    torch.bfloat16 --> F32, shape = {3072}
INFO:hf-to-gguf:blk.1.ffn_down.weight,     torch.bfloat16 --> Q8_0, shape = {8192, 3072}
INFO:hf-to-gguf:blk.1.ffn_up.weight,       torch.bfloat16 --> Q8_0, shape = {3072, 16384}
INFO:hf-to-gguf:blk.1.ffn_norm.weight,     torch.bfloat16 --> F32, shape = {3072}
INFO:hf-to-gguf:blk.1.attn_output.weight,  torch.bfloat16 --> Q8_0, shape = {3072, 3072}
INFO:hf-to-gguf:blk.1.attn_qkv.weight,     torch.bfloat16 --> Q8_0, shape = {3072, 9216}
INFO:hf-to-gguf:blk.10.attn_norm.weight,   torch.bfloat16 --> F32, shape = {3072}
```

Cargamos el modelo de chef bot en LM Studio.



Aquí está el modelo en formato gguf.



Prueba

Probamos el modelo haciéndole una pregunta sobre una receta, y como se puede ver, responde correctamente.