Carga del Modelo Base

Vamos a traer a Qwen2.5-Coder-14B a la memoria de tu RTX 3090, comprimido en 4-bits para que entre sin problemas.

In [17]:
```
import os       # Permite interactuar con el sistema operativo (rutas de carpetas, variables de entorno).
import gc       # "Garbage Collector": Se usa para liberar memoria RAM/VRAM manualmente si es necesario.
import json     # Para manipular archivos JSON (lectura de datasets o configuraciones).
import torch    # La librería principal de PyTorch para operaciones con tensores y uso de la GPU.
import shutil   # Útil para operaciones de archivos de alto nivel, como borrar o mover carpetas completas.
import subprocess # Permite ejecutar comandos de terminal (como git o pip) desde Python.

# Importa la clase Dataset de Hugging Face para estructurar los datos de entrenamiento.
from datasets import Dataset

# Unsloth: Librería optimizada para entrenar modelos más rápido y con menos memoria.
from unsloth import FastLanguageModel, is_bfloat16_supported

# Permite obtener información técnica de un modelo alojado en el Hugging Face Hub.
from huggingface_hub import model_info

# Facilita la aplicación de formatos de chat (como Llama-3 o Alpaca) a los datos.
from unsloth.chat_templates import get_chat_template

# El "Entrenador" (Trainer) especializado en Supervised Fine-Tuning (SFT).
from trl import SFTTrainer

# Define los hiperparámetros del entrenamiento (épocas, tasa de aprendizaje, pasos, etc.).
from transformers import TrainingArguments
```

In [2]:
```
# --- Configuración de parámetros iniciales ---

# Define la longitud máxima de tokens (contexto) que el modelo procesará.
# 2048 es estándar, pero Unsloth permite ampliarlo dinámicamente.
max_seq_length = 2048

# El tipo de datos para los pesos (None deja que Unsloth lo detecte automáticamente).
# Usualmente detectará float16 o bfloat16 según tu GPU.
dtype = None

# Activa la cuantización de 4 bits. Crucial para que un modelo de 14B
# quepa en GPUs de consumo (como una RTX 3060/4060 o superiores).
load_in_4bit = True

# El identificador del modelo en Hugging Face.
# Esta versión ya viene pre-cuantizada ("bnb-4bit") para ser ultra rápida.
model_name = "unsloth/Qwen2.5-Coder-14B-Instruct-bnb-4bit"

# --- Verificación de archivos locales (Caché) ---

# Construye la ruta donde Hugging Face suele guardar los modelos descargados.
# Transforma "usuario/modelo" en el formato de carpetas de caché del sistema.
cache_dir = os.path.expanduser(f"~/.cache/huggingface/hub/models--{'--'.join(model_name.split('/'))}")

# Comprueba si la carpeta del modelo ya existe en el disco duro.
if os.path.exists(cache_dir):
    # Si existe, nos avisa que no gastará internet descargándolo de nuevo.
    print(f"✅ Modelo encontrado en caché: {cache_dir}")
    print("🔄 Cargando modelo localmente...")
else:
    # Si no existe, nos advierte que iniciará una descarga pesada.
    print(f"⚠️ Modelo NO encontrado en caché. Se descargará en: {cache_dir}")

# --- Carga del Modelo y el Tokenizador ---

# Utiliza la función optimizada de Unsloth para cargar el modelo en la VRAM de la GPU.
# Retorna dos objetos:
# 1. model: El cerebro del IA.
# 2. tokenizer: El traductor que convierte texto en números (tokens).
model, tokenizer = FastLanguageModel.from_pretrained(
    model_name=model_name,
    max_seq_length=max_seq_length,
    dtype=dtype,
    load_in_4bit=load_in_4bit,
)

# Confirmación final de que el modelo está listo para usarse o entrenarse.
print("✅ Modelo cargado exitosamente en 4-bits.")
```

```
✅ Modelo encontrado en caché: /root/.cache/huggingface/hub/models--unsloth--Qwen2.5-Coder-14B-Instruct-bnb-4bit
🔄 Cargando modelo localmente...
==((====))==  Unsloth 2026.2.1: Fast Qwen2 patching. Transformers: 4.57.6.
   \\   /|    NVIDIA GeForce RTX 4090. Num GPUs = 1. Max memory: 22.152 GB. Platform: Linux.
O^O/ \_/ \    Torch: 2.10.0+cu128. CUDA: 8.9. CUDA Toolkit: 12.8. Triton: 3.6.0
\        /    Bfloat16 = TRUE. FA [Xformers = 0.0.34. FA2 = False]
 "-____-"     Free license: http://github.com/unslothai/unsloth
Unsloth: Fast downloading is enabled - ignore downloading bars which are red colored!


  Fetching 2 files: 100%|██████████| 2/2 [00:55<00:00, 27.51s/it]
  Loading checkpoint shards: 100%|██████████| 2/2 [00:01<00:00,  1.12it/s]


✅ Modelo cargado exitosamente en 4-bits.
```

Verificar que el modelo existe manualmente

Si ninguna de las anteriores funciona, verifica que puedes acceder al modelo:

In [4]:
```
# --- Verificación de metadatos en Hugging Face Hub ---

# Usamos un bloque "try-except" para manejar posibles errores de conexión o permisos.
try:
    # Llama a la API de Hugging Face para obtener la ficha técnica (info) del modelo.
    # Esto no descarga el modelo, solo consulta sus estadísticas y estado.
    info = model_info("unsloth/Qwen2.5-Coder-14B-Instruct-bnb-4bit")

    # Si la consulta es exitosa, imprime el ID confirmado del modelo.
    print(f"✅ Modelo encontrado: {info.id}")

    # Muestra el número total de descargas que ha tenido el modelo (popularidad).
    print(f"📩 Descargas: {info.downloads}")

# Si ocurre un error (ej. no hay internet, el modelo es privado o el nombre está mal escrito):
except Exception as e:
    # Atrapa el error y lo muestra en pantalla sin detener la ejecución del programa.
    print(f"❌ Error accediendo al modelo: {e}")
```

```
✅ Modelo encontrado: unsloth/Qwen2.5-Coder-14B-Instruct-bnb-4bit
📩 Descargas: 8706
```

Inyección de Adaptadores (LoRA)

Aquí definimos la arquitectura del fine-tuning. Solo vamos a entrenar una fracción del modelo (los adaptadores), lo que hace que el proceso sea rápido y eficiente.

In [3]:
```
# --- Configuración de PEFT (Parameter-Efficient Fine-Tuning) con LoRA ---

# Transforma el modelo base en un modelo PEFT (solo una parte es entrenable).
model = FastLanguageModel.get_peft_model(
    model,
    # 'r' (Rank): Define el tamaño de las matrices de bajo rango.
    # 16 es un equilibrio ideal entre precisión y ahorro de memoria.
    r = 16,

    # 'target_modules': Especifica en qué capas del modelo se inyectarán los adaptadores.
    # Estas capas (q, k, v, o, gate, up, down) cubren casi toda la atención y redes neuronales del modelo.
    target_modules = ["q_proj", "k_proj", "v_proj", "o_proj",
                      "gate_proj", "up_proj", "down_proj"],

    # 'lora_alpha': Escala el aprendizaje de los adaptadores.
    # Generalmente se recomienda que sea igual o el doble de 'r'.
    lora_alpha = 16,

    # 'lora_dropout': Probabilidad de desactivar neuronas al azar para evitar sobreajuste.
    # 0 es lo más eficiente para velocidad de entrenamiento en Unsloth.
    lora_dropout = 0,

    # 'bias': Define si se entrenan los sesgos. "none" es lo estándar para LoRA.
    bias = "none",

    # 'use_gradient_checkpointing': Técnica que libera memoria RAM de la GPU
    # guardando solo lo esencial. "unsloth" usa una versión optimizada que gasta un 30% menos.
    use_gradient_checkpointing = "unsloth",

    # Semilla aleatoria para que los resultados sean reproducibles (siempre den lo mismo).
    random_state = 3407,
)

# Mensaje de confirmación: el modelo ahora está listo para recibir datos de entrenamiento.
print("✅ Adaptadores LoRA configurados.")
```

```
Unsloth 2026.2.1 patched 48 layers with 48 QKV layers, 48 O layers and 48 MLP layers.
```

```
✅ Adaptadores LoRA configurados.
```

Preparación del Dataset
Le enseñamos al modelo a entender el formato de tu dataset.jsonl (ChatML: System, User, Assistant). Asumimos que dentro de tu JSONL, el arreglo de mensajes se llama messages o conversations.

Este código es fundamental: actúa como el traductor entre la forma en que tú guardaste la información en tu archivo `epics.jsonl` y la forma exacta en la que el modelo (Qwen2.5) necesita leerla para aprender.

En términos sencillos, el modelo no entiende "columnas" o "diccionarios"; solo entiende secuencias largas de texto con etiquetas especiales que le indican quién está hablando.

1. Aplicar la plantilla ChatML (`get_chat_template`)

Los modelos conversacionales como Qwen usan un formato llamado **ChatML** (Chat Markup Language). Esto significa que usan "etiquetas invisibles" para separar los mensajes, como `<|im_start|>user` y `<|im_end|>` .
Este paso configura tu tokenizador para que inyecte automáticamente estas etiquetas en el texto, ahorrándote el trabajo de escribirlas a mano.

2. Definir el System Prompt

Aquí le inyectamos la personalidad a tu PM Senior. Le estamos diciendo explícitamente cuál es su rol y, muy importante, le indicamos que **debe responder en formato JSON**. Esto ancla el comportamiento del modelo para que siempre actúe como un experto estructurado.

3. La función de formateo (`formatting_prompts_func`)

Este es el "motor" de la celda. Como tu archivo tiene una columna llamada `input` y otra llamada `output` , la función hace lo siguiente para cada fila de tu dataset:

* **Extrae los datos:** Toma el diccionario crudo de `input` y el de `output` .
* **Formatea a texto JSON legible (`json.dumps`):** Este paso es el truco vital. Como tu objetivo es que el modelo devuelva un JSON perfecto (como se ve en tus salidas con `epic_id` , `title` , `acceptance_criteria`), usamos `json.dumps(..., indent=2)` para transformar tus diccionarios en cadenas de texto con saltos de línea y tabulaciones perfectas. Así el modelo aprende a indentar como un humano.
* **Arma la conversación:** Crea una lista lógica con tres roles:
1. El `system` (las instrucciones base).
2. El `user` (tu `input` con el contexto y requerimientos).
3. El `assistant` (tu `output` con la Epic dorada).

* **Aplica la plantilla (`apply_chat_template`):** Pasa esa lista estructurada por el tokenizador para fusionarla en un único bloque de texto continuo con todas las etiquetas ChatML listas para el entrenamiento.

4. Cargar y procesar (`load_dataset` y `map`)

* Usa la librería de Hugging Face para cargar tu archivo `epics.jsonl` a la memoria RAM de tu RunPod.
* El comando `.map(..., batched=True)` pasa todo tu dataset por la función que explicamos arriba de forma simultánea y súper rápida.
El resultado final es que tu dataset ahora tiene una nueva columna llamada ***`text`*** *, que contiene la conversación perfectamente formateada. Esta columna* `text` *es la* única* que Qwen va a leer durante el entrenamiento.

In [5]:
```
# --- 1. Preparación del Formato de Conversación ---

# Configura el tokenizador para que use la estructura "ChatML" (<|im_start|>, <|im_end|>).
# Esto es vital para que el modelo sepa cuándo termina de hablar el usuario y empieza él.
tokenizer = get_chat_template(
    tokenizer,
    chat_template = "chatml",
)

# --- 2. Definición de la Identidad (System Prompt) ---

# Establecemos las "instrucciones de comportamiento" del modelo.
# Aquí le decimos que sea un Product Manager experto y que responda en JSON.
system_prompt = "Eres un Product Manager Senior experto. Tu tarea es analizar el contexto y los requerimientos proporcionados para
redactar Epics de software detalladas, estructuradas y precisas en formato JSON."

formatted_texts = []

# --- 3. Procesamiento Manual del Archivo de Datos ---

# Ruta donde tienes guardados tus ejemplos de entrenamiento (formato JSON Lines).
file_path = "../data/epics.jsonl"

with open(file_path, "r", encoding="utf-8") as f:
    for line in f:
        # Si la línea está vacía, nos la saltamos para evitar errores.
        if not line.strip(): continue

        # Convertimos la línea de texto (JSON) en un diccionario de Python.
        record = json.loads(line)

        # Convertimos los campos 'input' y 'output' en texto (strings) bien formateados.
        # 'ensure_ascii=False' permite tildes y ñ; 'indent=2' lo hace legible.
        user_content = json.dumps(record["input"], ensure_ascii=False, indent=2)
        assistant_content = json.dumps(record["output"], ensure_ascii=False, indent=2)

        # Creamos la estructura de la conversación (Mensaje de Sistema -> Usuario -> Asistente).
        convo = [
            {"role": "system", "content": system_prompt},
            {"role": "user", "content": user_content},
            {"role": "assistant", "content": assistant_content}
        ]

        # 'apply_chat_template' une todo lo anterior usando las etiquetas especiales de ChatML.
        # tokenize=False: Solo genera el texto plano por ahora.
        text = tokenizer.apply_chat_template(convo, tokenize=False, add_generation_prompt=False)

        # Guardamos el resultado en una lista de diccionarios con la clave "text".
        formatted_texts.append({"text": text})

# --- 4. Creación del Dataset Final ---

# Convertimos nuestra lista de Python en un objeto Dataset de HuggingFace.
# Este formato es el que el 'SFTTrainer' de Unsloth requiere para empezar a entrenar.
dataset = Dataset.from_list(formatted_texts)

# Mensajes de control para verificar que todo salió bien.
print(f"✅ ¡Dataset cargado y formateado! Total de ejemplos procesados: {len(dataset)}")

# Imprimimos los primeros 500 caracteres del primer ejemplo para ver las etiquetas <|im_start|>.
print("\n--- MUESTRA DEL PRIMER EJEMPLO FORMATEADO ---")
print(dataset[0]["text"][:500] + "...\n[CONTINÚA]")
```

```
Unsloth: Will map <|im_end|> to EOS = <|im_end|>.
```

```
✅ ¡Dataset cargado y formateado! Total de ejemplos procesados: 104

--- MUESTRA DEL PRIMER EJEMPLO FORMATEADO ---
<|im_start|>system
Eres un Product Manager Senior experto. Tu tarea es analizar el contexto y los requerimientos proporcionados para redactar
Epics de software detalladas, estructuradas y precisas en formato JSON.<|im_end|>
<|im_start|>user
{
  "context": "El proyecto inicia desde cero, sin ningún tipo de infraestructura en la nube. Es imperativo establecer una base
sólida, repetible y segura que permita el despliegue y la operación de todos los componentes subsecuentes de la plataforma.
La adop...
[CONTINÚA]
```

Fine-Tunning (ajusta) un modelo de lenguaje pre-entrenado con tus datos específicos usando **LoRA**

Componente	Función
`SFTTrainer`	Entrenador especializado para "Supervised Fine-Tuning" (ajuste supervisado)
`train_dataset`	Tus datos de entrenamiento (104 ejemplos conversacionales)
`max_seq_length`	Límite de tokens por ejemplo (2048)
`packing=False`	Respeta la estructura conversacional exacta de cada ejemplo
`per_device_train_batch_size=2`	Procesa 2 ejemplos simultáneamente en GPU
`gradient_accumulation_steps=4`	Simula un lote de 8 ejemplos (2×4) para ahorrar VRAM
`num_train_epochs=3`	Pasa 3 veces por todo el dataset (~39 pasos totales)
`learning_rate=2e-4`	Velocidad de aprendizaje estándar para LoRA
`adamw_8bit`	Optimizador comprimido que usa menos memoria
`fp16/bf16`	Precisión mixta para acelerar entrenamiento

Tiempo estimado: ~39 pasos × tiempo por paso (varía según GPU).

In [6]:

```
# --- 0. LIMPIEZA PROFUNDA DE GPU ---
# Libera cualquier residuo de memoria en la GPU para empezar desde cero.
torch.cuda.empty_cache()
# Fuerza al recolector de basura de Python a limpiar objetos no utilizados en la RAM.
gc.collect()

# --- 1. CARGA DEL MODELO ---
# Aumentamos la longitud de secuencia a 4096 para manejar Epics más largas.
max_seq_length = 4096
print("⏳ Cargando modelo base...")
model, tokenizer = FastLanguageModel.from_pretrained(
    model_name = "unsloth/Qwen2.5-Coder-14B-Instruct-bnb-4bit",
    max_seq_length = max_seq_length,
    dtype = None,
    load_in_4bit = True, # Cuantización para reducir el peso del modelo en VRAM.
)

# --- 2. ADAPTADORES LORA (MODO LIGERO) ---
print("🧠 Inyectando adaptadores LoRA (Modo Ligero)...")
model = FastLanguageModel.get_peft_model(
    model,
    r = 8, # Reducimos el rango (Rank). Menos parámetros entrenables = menos memoria.
    # Solo entrenamos las capas de atención (proyecciones Q, K, V, O).
    # Al quitar "gate_proj", "up_proj" y "down_proj", ahorramos mucha VRAM.
    target_modules = ["q_proj", "k_proj", "v_proj", "o_proj"],
    lora_alpha = 8,
    lora_dropout = 0,
    bias = "none",
    use_gradient_checkpointing = "unsloth",
    random_state = 3407,
)

# --- 3. PREPARAR DATASET ---
# (Este bloque aplica la plantilla ChatML y limpia el JSONL como vimos anteriormente)
print("📊 Formateando el dataset...")
tokenizer = get_chat_template(tokenizer, chat_template = "chatml")
system_prompt = "Eres un Product Manager Senior experto..."
formatted_texts = []

with open("../data/epics.jsonl", "r", encoding="utf-8") as f:
    for line in f:
        if not line.strip(): continue
        record = json.loads(line)
        user_content = json.dumps(record["input"], ensure_ascii=False, indent=2)
        assistant_content = json.dumps(record["output"], ensure_ascii=False, indent=2)
        convo = [
            {"role": "system", "content": system_prompt},
            {"role": "user", "content": user_content},
            {"role": "assistant", "content": assistant_content}
        ]
        text = tokenizer.apply_chat_template(convo, tokenize=False, add_generation_prompt=False)
        formatted_texts.append({"text": text})

dataset = Dataset.from_list(formatted_texts)

# --- 4. ENTRENAMIENTO BLINDADO CONTRA OOM ---
print("🚀 ¡INICIANDO FINE-TUNING EXTREMO!")
trainer = SFTTrainer(
    model = model,
    tokenizer = tokenizer,
    train_dataset = dataset,
    dataset_text_field = "text",
    max_seq_length = max_seq_length,
    dataset_num_proc = 2, # Usa 2 procesos para cargar datos más rápido.
    packing = False,      # No empaqueta secuencias cortas (más lento pero más estable).
    args = TrainingArguments(
        per_device_train_batch_size = 1, # Procesa 1 ejemplo a la vez para no saturar la GPU.
        gradient_accumulation_steps = 8, # Actualiza pesos cada 8 pasos (Batch efectivo = 8).
        warmup_steps = 5,                # Sube la intensidad del aprendizaje gradualmente.
        num_train_epochs = 3,            # El modelo verá el dataset completo 3 veces.
        learning_rate = 2e-4,            # Velocidad de aprendizaje (estándar para LoRA).
        # Selecciona automáticamente entre fp16 o bf16 según la potencia de tu GPU.
        fp16 = not is_bfloat16_supported(),
        bf16 = is_bfloat16_supported(),
        logging_steps = 1,               # Muestra el progreso en cada paso.
        # "paged_adamw_8bit": Si la GPU se queda sin memoria, usa la RAM del sistema como "colchón".
        optim = "paged_adamw_8bit",
        weight_decay = 0.01,
        lr_scheduler_type = "linear",
        seed = 3407,
        output_dir = "outputs",          # Carpeta donde se guardarán los checkpoints.
    ),
)

# Inicia el proceso de entrenamiento y guarda las estadísticas finales.
trainer_stats = trainer.train()
print(f"✅ ¡Entrenamiento completado en {trainer_stats.metrics['train_runtime']} segundos!")
```

```
⏳ Cargando modelo base...
==((====))==  Unsloth 2026.2.1: Fast Qwen2 patching. Transformers: 4.57.6.
   \\   /|    NVIDIA GeForce RTX 4090. Num GPUs = 1. Max memory: 22.152 GB. Platform: Linux.
O^O/ \_/ \    Torch: 2.10.0+cu128. CUDA: 8.9. CUDA Toolkit: 12.8. Triton: 3.6.0
\        /    Bfloat16 = TRUE. FA [Xformers = 0.0.34. FA2 = False]
 "-____-"     Free license: http://github.com/unslothai/unsloth
Unsloth: Fast downloading is enabled - ignore downloading bars which are red colored!


 Loading checkpoint shards: 100%|██████████| 2/2 [00:02<00:00,  1.24s/it]


🧠 Inyectando adaptadores LoRA (Modo Ligero)...


 Not an error, but Unsloth cannot patch MLP layers with our manual autograd engine since either LoRA adapters
 are not enabled or a bias term (like in Qwen) is used.
 Unsloth 2026.2.1 patched 48 layers with 48 QKV layers, 48 O layers and 0 MLP layers.
```

📊 Formateando el dataset...
🚀 ¡INICIANDO FINE-TUNING EXTREMO!

```
Unsloth: Tokenizing ["text"] (num_proc=64): 100%|██████████| 104/104 [00:13<00:00,  7.70 examples/s]
==((====))==  Unsloth - 2x faster free finetuning | Num GPUs used = 1
   \\   /|    Num examples = 104 | Num Epochs = 3 | Total steps = 39
O^O/ \_/ \    Batch size per device = 1 | Gradient accumulation steps = 8
\        /    Data Parallel GPUs = 1 | Total batch size (1 x 8 x 1) = 8
 "-____-"     Trainable parameters = 12,582,912 of 14,782,616,576 (0.09% trained)
```

[39/39 02:37, Epoch 3/3]

Step	Training Loss
1	1.863400
2	1.693400
3	1.883100
4	1.722500
5	1.803800
6	1.815500
7	1.813900
8	1.768000
9	1.707100
10	1.632500
11	1.705800
12	1.545300
13	1.571100
14	1.492200
15	1.603400
16	1.507500
17	1.511300
18	1.462900
19	1.482600
20	1.372300
21	1.440500
22	1.491500
23	1.445000
24	1.522100
25	1.361400
26	1.500200
27	1.311200
28	1.435300
29	1.341400
30	1.416900
31	1.371400
32	1.337000
33	1.313900
34	1.331500
35	1.401100
36	1.354400
37	1.405900
38	1.408500
39	1.316600

✅ ¡Entrenamiento completado en 165.8221 segundos!

In [8]:
```
# 'apt-get update -y': Actualiza la lista de paquetes disponibles en los repositorios.
# El flag '-y' responde automáticamente "sí" a las confirmaciones para que no se detenga.
!apt-get update -y && \

# 'apt-get install -y': Comando para instalar nuevos paquetes.
# Se instalan 3 componentes fundamentales para compilar software:
# 1. cmake: Herramienta avanzada para gestionar el proceso de compilación (indispensable para muchas librerías de IA).
# 2. build-essential: Un paquete que incluye el compilador GCC, G++ y herramientas básicas para crear software desde el código
fuente.
# 3. libcurl4-openssl-dev: Librería necesaria para que las aplicaciones puedan realizar transferencias de red (como descargar
modelos o comunicarse con APIs) de forma segura.
apt-get install -y cmake build-essential libcurl4-openssl-dev
```

```
Hit:1 http://archive.ubuntu.com/ubuntu jammy InRelease
Hit:2 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86_64  InRelease
Get:3 http://archive.ubuntu.com/ubuntu jammy-updates InRelease [128 kB]
Hit:4 https://ppa.launchpadcontent.net/deadsnakes/ppa/ubuntu jammy InRelease
Get:5 http://security.ubuntu.com/ubuntu jammy-security InRelease [129 kB]
Get:6 http://archive.ubuntu.com/ubuntu jammy-backports InRelease [127 kB]
Fetched 384 kB in 1s (433 kB/s)
Reading package lists... Done
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
build-essential is already the newest version (12.9ubuntu3).
build-essential set to manually installed.
cmake is already the newest version (3.22.1-1ubuntu1.22.04.2).
libcurl4-openssl-dev is already the newest version (7.81.0-1ubuntu1.21).
0 upgraded, 0 newly installed, 0 to remove and 134 not upgraded.
```

In [14]:
```
# '!': Ejecuta el comando en la consola del sistema.
# 'poetry add': Instala las librerías y las registra en tu archivo de proyecto.

!poetry add \
    # 'gguf': Necesario para escribir y leer el formato de archivo final (.gguf).
    gguf \

    # 'protobuf': El sistema de serialización de datos que usa Google y Hugging Face.
    protobuf \

    # 'sentencepiece': ¡La pieza clave! Es la librería que maneja la tokenización
    # de modelos como Qwen, Llama y Mistral. Sin esto, el modelo no puede
    # descomponer las palabras en unidades que entienda (tokens).
    sentencepiece
```

```
The following packages are already present in the pyproject.toml and will be skipped:

  - [36mgguf [39m
  - [36mprotobuf [39m

If you want to update it to the latest compatible version, you can use `poetry update package`.
If you prefer to upgrade it to the latest available version, you can use `poetry add package@latest`.

Using version  [39;1m^0.2.1 [39;22m for  [36msentencepiece [39m

 [34mUpdating dependencies [39m
 [2K [34mResolving dependencies... [39m  [39;2m(0.6s) [39;22m

No dependencies to install or update

 [34mWriting lock file [39m
```

In [11]:
```
# '!': Indica que el comando se ejecuta en la consola del sistema (Shell).
# 'rm': Es el comando para "remover" (borrar) archivos o directorios.

!rm -rf \
    # '-r' (recursive): Permite borrar carpetas y todo su contenido interno (subcarpetas y archivos).
    # '-f' (force): Fuerza el borrado ignorando archivos inexistentes y sin pedir confirmación al usuario.
    -rf \

    # 'llama.cpp': El nombre de la carpeta específica que se desea eliminar.
    llama.cpp
```

In [16]:

```
# --- Verificación y Setup de llama.cpp ---

# Definimos la ruta local donde queremos que viva el repositorio llama.cpp.
LLAMA_CPP_DIR = "./llama.cpp"

# Comprobamos si la carpeta ya existe para no descargarla dos veces.
if not os.path.exists(LLAMA_CPP_DIR):
    print("📥 Descargando llama.cpp...")
    # 'git clone': Descarga el código fuente oficial de llama.cpp.
    subprocess.run([
        "git", "clone",
        "https://github.com/ggerganov/llama.cpp.git",
        LLAMA_CPP_DIR
    ], check=True)
    print("✅ llama.cpp descargado")

    # Instalamos las librerías de Python necesarias para que los scripts de conversión funcionen.
    print("📦 Instalando dependencias...")
    subprocess.run([
        "pip", "install", "-r",
        f"{LLAMA_CPP_DIR}/requirements.txt"
    ], check=True)
else:
    # Si la carpeta ya está ahí, simplemente lo confirmamos.
    print("✅ llama.cpp ya existe")

# --- Proceso de Exportación ---

print("📦 Iniciando fusión y exportación a GGUF...")
print("⏳ Esto puede tomar 10-20 minutos...")

try:
    # Intento A: Usar la función integrada de Unsloth.
    model.save_pretrained_gguf(
        "MiPM_Senior",
        tokenizer,
        quantization_method="q4_k_m", # Cuantización de 4 bits (calidad media-alta).
        # Indicamos explícitamente dónde está la herramienta de conversión.
        converter_location=LLAMA_CPP_DIR,
    )
    print("✅ ¡PROCESO COMPLETADO!")
    print("📁 Archivo generado: MiPM_Senior_q4_k_m.gguf")

except Exception as e:
    # Intento B (Fallback): Si lo anterior falla (por errores de memoria o librerías),
    # hacemos el proceso en dos pasos manuales.
    print(f"⚠️ Error con método automático: {e}")
    print("🔄 Intentando método manual...")

    # 1. Primero, fusionamos el modelo LoRA con el base y lo guardamos como un modelo normal de Hugging Face.
    # 'merged_16bit' asegura que no haya pérdida de calidad en esta fase.
    model.save_pretrained_merged(
        "MiPM_Senior_HF",
        tokenizer,
        save_method="merged_16bit",
    )

    print("✅ Modelo guardado en formato HF")
    # 2. Instrucciones para que el usuario ejecute la conversión final desde la consola.
    print("📝 Para convertir a GGUF manualmente, ejecuta en terminal:")
    print(f"""
    cd {LLAMA_CPP_DIR}
    python convert_hf_to_gguf.py ../MiPM_Senior_HF --outfile ../MiPM_Senior.gguf --outtype q4_k_m
    """)
```

```
✅ llama.cpp ya existe
📦 Iniciando fusión y exportación a GGUF...
⏳ Esto puede tomar 10-20 minutos...
⚠️ Error con método automático: unsloth_save_pretrained_gguf() got an unexpected keyword argument 'converter_location'
🔄 Intentando método manual...
Found HuggingFace hub cache directory: /root/.cache/huggingface/hub
```

```
Fetching 1 files: 100%|██████████| 1/1 [00:00<00:00,  2.01it/s]

Checking cache directory for required files...
Cache check failed: model-00001-of-00006.safetensors not found in local cache.
Not all required files found in cache. Will proceed with downloading.
Checking cache directory for required files...
Cache check failed: tokenizer.model not found in local cache.
Not all required files found in cache. Will proceed with downloading.

Unsloth: Preparing safetensor model files: 100%|██████████| 6/6 [02:28<00:00, 24.71s/it]

Note: tokenizer.model not found (this is OK for non-SentencePiece models)

Unsloth: Merging weights into 16bit: 100%|██████████| 6/6 [01:10<00:00, 11.70s/it]

Unsloth: Merge process complete. Saved to `/workspace/code/MiPM_Senior_HF`
✅ Modelo guardado en formato HF
📝 Para convertir a GGUF manualmente, ejecuta en terminal:

    cd ./llama.cpp
    python convert_hf_to_gguf.py ../MiPM_Senior_HF --outfile ../MiPM_Senior.gguf --outtype q4_k_m
```

In [21]:
```
# Instalar herramientas de compilación esenciales
!apt-get update && apt-get install -y build-essential gcc g++ make

# Verificar espacio en disco en /workspace
print("\n--- ESPACIO EN DISCO ---")
!df -h /workspace
```

```
Hit:1 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86_64  InRelease
Hit:2 https://ppa.launchpadcontent.net/deadsnakes/ppa/ubuntu jammy InRelease
Hit:3 http://archive.ubuntu.com/ubuntu jammy InRelease
Hit:4 http://archive.ubuntu.com/ubuntu jammy-updates InRelease
Hit:5 http://security.ubuntu.com/ubuntu jammy-security InRelease
Hit:6 http://archive.ubuntu.com/ubuntu jammy-backports InRelease
Reading package lists... Done
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
build-essential is already the newest version (12.9ubuntu3).
g++ is already the newest version (4:11.2.0-1ubuntu1).
g++ set to manually installed.
gcc is already the newest version (4:11.2.0-1ubuntu1).
gcc set to manually installed.
make is already the newest version (4.3-4.1build1).
make set to manually installed.
0 upgraded, 0 newly installed, 0 to remove and 134 not upgraded.

--- ESPACIO EN DISCO ---
Filesystem                    Size  Used Avail Use% Mounted on
mfs#eur-no-1.runpod.net:9421 1006T  591T  416T  59% /workspace
```

In [ ]:

```
import os
import subprocess
import shutil

def convert_and_quantize_manual(model_path, output_final_gguf):
    print(f"🚀 Iniciando proceso para: {model_path}")

    # 1. Clonar llama.cpp si no existe
    if not os.path.exists("llama.cpp"):
        print("📥 Clonando llama.cpp...")
        subprocess.run(["git", "clone", "https://github.com/ggerganov/llama.cpp"], check=True)

    # 2. Instalar dependencias dentro de Poetry
    print("📦 Asegurando dependencias en entorno Poetry...")
    subprocess.run(["poetry", "run", "pip", "install", "gguf", "sentencepiece", "protobuf"], check=True)
    subprocess.run(["poetry", "run", "pip", "install", "-r", "llama.cpp/requirements.txt"], check=True)

    # Rutas
    temp_f16_gguf = "/workspace/model_temp_f16.gguf"
    build_dir = "llama.cpp/build"

    try:
        # PASO 1: Conversión HF -> GGUF F16
        print("⚙️ PASO 1: Convirtiendo HF a GGUF F16...")
        conv_script = "llama.cpp/convert_hf_to_gguf.py"
        conv_cmd = [
            "poetry", "run", "python", conv_script,
            model_path,
            "--outfile", temp_f16_gguf,
            "--outtype", "f16"
        ]
        subprocess.run(conv_cmd, check=True)
        print("✅ Conversión a F16 completada.")

        # PASO 2: Compilar usando CMake (Nuevo sistema)
        print("🛠️ PASO 2: Compilando llama-quantize con CMake...")
        os.makedirs(build_dir, exist_ok=True)

        # Configurar el build
        subprocess.run(["cmake", "-B", build_dir, "-S", "llama.cpp", "-DGGML_CUDA=ON"], check=True)

        # Compilar solo el binario de cuantización (específicamente el target 'llama-quantize')
        subprocess.run(["cmake", "--build", build_dir, "--config", "Release", "--target", "llama-quantize", "-j"], check=True)

        # PASO 3: Cuantización de F16 a Q4_K_M
        print(f"💎 PASO 3: Cuantizando a q4_k_m -> {output_final_gguf}...")

        # La ruta del binario ahora está dentro de build/bin/
        quant_binary = os.path.join(build_dir, "bin", "llama-quantize")

        quant_cmd = [
            quant_binary,
            temp_f16_gguf,
            output_final_gguf,
            "q4_k_m"
        ]
        subprocess.run(quant_cmd, check=True)

        # Limpieza
        print("🧹 Limpiando archivos temporales...")
        if os.path.exists(temp_f16_gguf):
            os.remove(temp_f16_gguf)

        print(f"🎊 ¡PROCESO EXITOSO! Descarga tu modelo aquí: {output_final_gguf}")

    except subprocess.CalledProcessError as e:
        print(f"❌ Error en el comando: {e.cmd}")
        print(f"Código de salida: {e.returncode}")

# --- EJECUCIÓN ---
input_hf_path = "/workspace/code/MiPM_Senior_HF"
output_gguf_path = "/workspace/MiPM_Senior_Expert.gguf"

convert_and_quantize_manual(input_hf_path, output_gguf_path)
```

```
🚀 Iniciando proceso para: /workspace/code/MiPM_Senior_HF
📦 Asegurando dependencias en entorno Poetry...
Requirement already satisfied: gguf in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (0.17.1)
Requirement already satisfied: sentencepiece in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (0.2.1)
Requirement already satisfied: protobuf in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (4.25.8)
Requirement already satisfied: numpy>=1.17 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from gguf) (1.26.4)
Requirement already satisfied: pyyaml>=5.1 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from gguf) (6.0.3)
Requirement already satisfied: tqdm>=4.27 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from gguf) (4.67.3)
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
Requirement already satisfied: numpy~=1.26.4 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from -r /workspace/code/llama.cpp/requirements/requirements-convert_legacy_llama.txt
(line 1)) (1.26.4)
Requirement already satisfied: sentencepiece<0.3.0,>=0.1.98 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from -r /workspace/code/llama.cpp/requirements/requirements-convert_legacy_llama.txt
(line 2)) (0.2.1)
Requirement already satisfied: transformers<5.0.0,>=4.57.1 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from -r /workspace/code/llama.cpp/requirements/requirements-convert_legacy_llama.txt
(line 4)) (4.57.6)
Requirement already satisfied: gguf>=0.1.0 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from -r /workspace/code/llama.cpp/requirements/requirements-convert_legacy_llama.txt
(line 6)) (0.17.1)
```

```
Requirement already satisfied: protobuf<5.0.0,>=4.21.0 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from -r /workspace/code/llama.cpp/requirements/requirements-convert_legacy_llama.txt
(line 7)) (4.25.8)
Requirement already satisfied: torch~=2.6.0 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from -r /workspace/code/llama.cpp/requirements/requirements-convert_hf_to_gguf.txt (line
5)) (2.6.0+cpu)
Requirement already satisfied: aiohttp~=3.9.3 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from -r /workspace/code/llama.cpp/requirements/requirements-tool_bench.txt (line 1))
(3.9.5)
Requirement already satisfied: pytest~=8.3.3 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from -r /workspace/code/llama.cpp/requirements/requirements-tool_bench.txt (line 2))
(8.3.5)
Requirement already satisfied: huggingface_hub<1.0,>=0.34.0 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from -r /workspace/code/llama.cpp/requirements/requirements-tool_bench.txt (line 3))
(0.36.2)
Requirement already satisfied: matplotlib~=3.10.0 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from -r /workspace/code/llama.cpp/requirements/requirements-tool_bench.txt (line 4))
(3.10.8)
Requirement already satisfied: openai~=2.14.0 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from -r /workspace/code/llama.cpp/requirements/requirements-tool_bench.txt (line 6))
(2.14.0)
Requirement already satisfied: pandas~=2.2.3 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from -r /workspace/code/llama.cpp/requirements/requirements-tool_bench.txt (line 7))
(2.2.3)
Requirement already satisfied: prometheus-client~=0.20.0 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from -r /workspace/code/llama.cpp/requirements/requirements-tool_bench.txt (line 8))
(0.20.0)
Requirement already satisfied: requests~=2.32.3 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from -r /workspace/code/llama.cpp/requirements/requirements-tool_bench.txt (line 9))
(2.32.5)
Requirement already satisfied: wget~=3.2 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from -r /workspace/code/llama.cpp/requirements/requirements-tool_bench.txt (line 10))
(3.2)
Requirement already satisfied: typer~=0.15.1 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from -r /workspace/code/llama.cpp/requirements/requirements-tool_bench.txt (line 11))
(0.15.4)
Requirement already satisfied: seaborn~=0.13.2 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from -r /workspace/code/llama.cpp/requirements/requirements-tool_bench.txt (line 12))
(0.13.2)
Requirement already satisfied: filelock in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from transformers<5.0.0,>=4.57.1->-r
/workspace/code/llama.cpp/requirements/requirements-convert_legacy_llama.txt (line 4)) (3.24.3)
Requirement already satisfied: packaging>=20.0 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from transformers<5.0.0,>=4.57.1->-r
/workspace/code/llama.cpp/requirements/requirements-convert_legacy_llama.txt (line 4)) (26.0)
Requirement already satisfied: pyyaml>=5.1 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from transformers<5.0.0,>=4.57.1->-r
/workspace/code/llama.cpp/requirements/requirements-convert_legacy_llama.txt (line 4)) (6.0.3)
Requirement already satisfied: regex!=2019.12.17 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from transformers<5.0.0,>=4.57.1->-r
/workspace/code/llama.cpp/requirements/requirements-convert_legacy_llama.txt (line 4)) (2026.2.19)
Requirement already satisfied: tokenizers<=0.23.0,>=0.22.0 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from transformers<5.0.0,>=4.57.1->-r
/workspace/code/llama.cpp/requirements/requirements-convert_legacy_llama.txt (line 4)) (0.22.2)
Requirement already satisfied: safetensors>=0.4.3 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from transformers<5.0.0,>=4.57.1->-r
/workspace/code/llama.cpp/requirements/requirements-convert_legacy_llama.txt (line 4)) (0.7.0)
Requirement already satisfied: tqdm>=4.27 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from transformers<5.0.0,>=4.57.1->-r
/workspace/code/llama.cpp/requirements/requirements-convert_legacy_llama.txt (line 4)) (4.67.3)
Requirement already satisfied: typing-extensions>=4.10.0 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from torch~=2.6.0->-r /workspace/code/llama.cpp/requirements/requirements-
convert_hf_to_gguf.txt (line 5)) (4.15.0)
Requirement already satisfied: networkx in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from torch~=2.6.0->-r /workspace/code/llama.cpp/requirements/requirements-
convert_hf_to_gguf.txt (line 5)) (3.6.1)
Requirement already satisfied: jinja2 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from torch~=2.6.0->-r /workspace/code/llama.cpp/requirements/requirements-
convert_hf_to_gguf.txt (line 5)) (3.1.6)
Requirement already satisfied: fsspec in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from torch~=2.6.0->-r /workspace/code/llama.cpp/requirements/requirements-
convert_hf_to_gguf.txt (line 5)) (2025.9.0)
Requirement already satisfied: sympy==1.13.1 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from torch~=2.6.0->-r /workspace/code/llama.cpp/requirements/requirements-
convert_hf_to_gguf.txt (line 5)) (1.13.1)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from sympy==1.13.1->torch~=2.6.0->-r
/workspace/code/llama.cpp/requirements/requirements-convert_hf_to_gguf.txt (line 5)) (1.3.0)
Requirement already satisfied: aiosignal>=1.1.2 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from aiohttp~=3.9.3->-r /workspace/code/llama.cpp/requirements/requirements-
tool_bench.txt (line 1)) (1.4.0)
Requirement already satisfied: attrs>=17.3.0 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from aiohttp~=3.9.3->-r /workspace/code/llama.cpp/requirements/requirements-
tool_bench.txt (line 1)) (25.4.0)
Requirement already satisfied: frozenlist>=1.1.1 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from aiohttp~=3.9.3->-r /workspace/code/llama.cpp/requirements/requirements-
tool_bench.txt (line 1)) (1.8.0)
Requirement already satisfied: multidict<7.0,>=4.5 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from aiohttp~=3.9.3->-r /workspace/code/llama.cpp/requirements/requirements-
tool_bench.txt (line 1)) (6.7.1)
Requirement already satisfied: yarl<2.0,>=1.0 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from aiohttp~=3.9.3->-r /workspace/code/llama.cpp/requirements/requirements-
tool_bench.txt (line 1)) (1.22.0)
Requirement already satisfied: iniconfig in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from pytest~=8.3.3->-r /workspace/code/llama.cpp/requirements/requirements-
tool_bench.txt (line 2)) (2.3.0)
Requirement already satisfied: pluggy<2,>=1.5 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from pytest~=8.3.3->-r /workspace/code/llama.cpp/requirements/requirements-
tool_bench.txt (line 2)) (1.6.0)
Requirement already satisfied: hf-xet<2.0.0,>=1.1.3 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from huggingface_hub<1.0,>=0.34.0->-r
/workspace/code/llama.cpp/requirements/requirements-tool_bench.txt (line 3)) (1.2.0)
Requirement already satisfied: contourpy>=1.0.1 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from matplotlib~=3.10.0->-r /workspace/code/llama.cpp/requirements/requirements-
tool_bench.txt (line 4)) (1.3.3)
Requirement already satisfied: cycler>=0.10 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from matplotlib~=3.10.0->-r /workspace/code/llama.cpp/requirements/requirements-
```

```
tool_bench.txt (line 4)) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from matplotlib~=3.10.0->-r /workspace/code/llama.cpp/requirements/requirements-
tool_bench.txt (line 4)) (4.61.1)
Requirement already satisfied: kiwisolver>=1.3.1 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from matplotlib~=3.10.0->-r /workspace/code/llama.cpp/requirements/requirements-
tool_bench.txt (line 4)) (1.4.9)
Requirement already satisfied: pillow>=8 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from matplotlib~=3.10.0->-r /workspace/code/llama.cpp/requirements/requirements-
tool_bench.txt (line 4)) (12.1.1)
Requirement already satisfied: pyparsing>=3 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from matplotlib~=3.10.0->-r /workspace/code/llama.cpp/requirements/requirements-
tool_bench.txt (line 4)) (3.3.2)
Requirement already satisfied: python-dateutil>=2.7 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from matplotlib~=3.10.0->-r /workspace/code/llama.cpp/requirements/requirements-
tool_bench.txt (line 4)) (2.9.0.post0)
Requirement already satisfied: anyio<5,>=3.5.0 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from openai~=2.14.0->-r /workspace/code/llama.cpp/requirements/requirements-
tool_bench.txt (line 6)) (4.12.1)
Requirement already satisfied: distro<2,>=1.7.0 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from openai~=2.14.0->-r /workspace/code/llama.cpp/requirements/requirements-
tool_bench.txt (line 6)) (1.9.0)
Requirement already satisfied: httpx<1,>=0.23.0 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from openai~=2.14.0->-r /workspace/code/llama.cpp/requirements/requirements-
tool_bench.txt (line 6)) (0.28.1)
Requirement already satisfied: jiter<1,>=0.10.0 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from openai~=2.14.0->-r /workspace/code/llama.cpp/requirements/requirements-
tool_bench.txt (line 6)) (0.13.0)
Requirement already satisfied: pydantic<3,>=1.9.0 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from openai~=2.14.0->-r /workspace/code/llama.cpp/requirements/requirements-
tool_bench.txt (line 6)) (2.12.5)
Requirement already satisfied: sniffio in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from openai~=2.14.0->-r /workspace/code/llama.cpp/requirements/requirements-
tool_bench.txt (line 6)) (1.3.1)
Requirement already satisfied: pytz>=2020.1 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from pandas~=2.2.3->-r /workspace/code/llama.cpp/requirements/requirements-
tool_bench.txt (line 7)) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from pandas~=2.2.3->-r /workspace/code/llama.cpp/requirements/requirements-
tool_bench.txt (line 7)) (2025.3)
Requirement already satisfied: charset_normalizer<4,>=2 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from requests~=2.32.3->-r /workspace/code/llama.cpp/requirements/requirements-
tool_bench.txt (line 9)) (3.4.4)
Requirement already satisfied: idna<4,>=2.5 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from requests~=2.32.3->-r /workspace/code/llama.cpp/requirements/requirements-
tool_bench.txt (line 9)) (3.11)
Requirement already satisfied: urllib3<3,>=1.21.1 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from requests~=2.32.3->-r /workspace/code/llama.cpp/requirements/requirements-
tool_bench.txt (line 9)) (2.6.3)
Requirement already satisfied: certifi>=2017.4.17 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from requests~=2.32.3->-r /workspace/code/llama.cpp/requirements/requirements-
tool_bench.txt (line 9)) (2026.1.4)
Requirement already satisfied: click<8.2,>=8.0.0 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from typer~=0.15.1->-r /workspace/code/llama.cpp/requirements/requirements-
tool_bench.txt (line 11)) (8.1.8)
Requirement already satisfied: shellingham>=1.3.0 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from typer~=0.15.1->-r /workspace/code/llama.cpp/requirements/requirements-
tool_bench.txt (line 11)) (1.5.4)
Requirement already satisfied: rich>=10.11.0 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from typer~=0.15.1->-r /workspace/code/llama.cpp/requirements/requirements-
tool_bench.txt (line 11)) (14.3.3)
Requirement already satisfied: httpcore==1.* in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from httpx<1,>=0.23.0->openai~=2.14.0->-r
/workspace/code/llama.cpp/requirements/requirements-tool_bench.txt (line 6)) (1.0.9)
Requirement already satisfied: h11>=0.16 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from httpcore==1.*->httpx<1,>=0.23.0->openai~=2.14.0->-r
/workspace/code/llama.cpp/requirements/requirements-tool_bench.txt (line 6)) (0.16.0)
Requirement already satisfied: annotated-types>=0.6.0 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from pydantic<3,>=1.9.0->openai~=2.14.0->-r
/workspace/code/llama.cpp/requirements/requirements-tool_bench.txt (line 6)) (0.7.0)
Requirement already satisfied: pydantic-core==2.41.5 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from pydantic<3,>=1.9.0->openai~=2.14.0->-r
/workspace/code/llama.cpp/requirements/requirements-tool_bench.txt (line 6)) (2.41.5)
Requirement already satisfied: typing-inspection>=0.4.2 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from pydantic<3,>=1.9.0->openai~=2.14.0->-r
/workspace/code/llama.cpp/requirements/requirements-tool_bench.txt (line 6)) (0.4.2)
Requirement already satisfied: propcache>=0.2.1 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from yarl<2.0,>=1.0->aiohttp~=3.9.3->-r
/workspace/code/llama.cpp/requirements/requirements-tool_bench.txt (line 1)) (0.4.1)
Requirement already satisfied: six>=1.5 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from python-dateutil>=2.7->matplotlib~=3.10.0->-r
/workspace/code/llama.cpp/requirements/requirements-tool_bench.txt (line 4)) (1.17.0)
Requirement already satisfied: markdown-it-py>=2.2.0 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from rich>=10.11.0->typer~=0.15.1->-r
/workspace/code/llama.cpp/requirements/requirements-tool_bench.txt (line 11)) (4.0.0)
Requirement already satisfied: pygments<3.0.0,>=2.13.0 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from rich>=10.11.0->typer~=0.15.1->-r
/workspace/code/llama.cpp/requirements/requirements-tool_bench.txt (line 11)) (2.19.2)
Requirement already satisfied: mdurl~=0.1 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from markdown-it-py>=2.2.0->rich>=10.11.0->typer~=0.15.1->-r
/workspace/code/llama.cpp/requirements/requirements-tool_bench.txt (line 11)) (0.1.2)
Requirement already satisfied: MarkupSafe>=2.0 in /root/.cache/pypoetry/virtualenvs/per-training-model-xS3fZVNL-
py3.11/lib/python3.11/site-packages (from jinja2->torch~=2.6.0->-r /workspace/code/llama.cpp/requirements/requirements-
convert_hf_to_gguf.txt (line 5)) (3.0.3)
⚙ PASO 1: Convirtiendo HF a GGUF F16...

INFO:hf-to-gguf:Loading model: MiPM_Senior_HF
INFO:hf-to-gguf:Model architecture: Qwen2ForCausalLM
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-00006.safetensors'
INFO:hf-to-gguf:gguf: indexing model part 'model-00002-of-00006.safetensors'
INFO:hf-to-gguf:gguf: indexing model part 'model-00003-of-00006.safetensors'
INFO:hf-to-gguf:gguf: indexing model part 'model-00004-of-00006.safetensors'
INFO:hf-to-gguf:gguf: indexing model part 'model-00005-of-00006.safetensors'
INFO:hf-to-gguf:gguf: indexing model part 'model-00006-of-00006.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 --> F16, shape = {5120, 152064}
INFO:hf-to-gguf:blk.0.attn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.0.ffn_down.weight,     torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.0.ffn_gate.weight,     torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.0.ffn_up.weight,       torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.0.ffn_norm.weight,     torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.0.attn_k.bias,         torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.0.attn_k.weight,       torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.0.attn_output.weight,  torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.0.attn_q.bias,         torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.0.attn_q.weight,       torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.0.attn_v.bias,         torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.0.attn_v.weight,       torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.1.attn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.1.ffn_down.weight,     torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.1.ffn_gate.weight,     torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.1.ffn_up.weight,       torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.1.ffn_norm.weight,     torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.1.attn_k.bias,         torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.1.attn_k.weight,       torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.1.attn_output.weight,  torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.1.attn_q.bias,         torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.1.attn_q.weight,       torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.1.attn_v.bias,         torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.1.attn_v.weight,       torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.2.attn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.2.ffn_down.weight,     torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.2.ffn_gate.weight,     torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.2.ffn_up.weight,       torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.2.ffn_norm.weight,     torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.2.attn_k.bias,         torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.2.attn_k.weight,       torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.2.attn_output.weight,  torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.2.attn_q.bias,         torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.2.attn_q.weight,       torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.2.attn_v.bias,         torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.2.attn_v.weight,       torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.3.attn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.3.ffn_down.weight,     torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.3.ffn_gate.weight,     torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.3.ffn_up.weight,       torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.3.ffn_norm.weight,     torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.3.attn_k.bias,         torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.3.attn_k.weight,       torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.3.attn_output.weight,  torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.3.attn_q.bias,         torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.3.attn_q.weight,       torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.3.attn_v.bias,         torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.3.attn_v.weight,       torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.4.attn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.4.ffn_down.weight,     torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.4.ffn_gate.weight,     torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.4.ffn_up.weight,       torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.4.ffn_norm.weight,     torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.4.attn_k.bias,         torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.4.attn_k.weight,       torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.4.attn_output.weight,  torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.4.attn_q.bias,         torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.4.attn_q.weight,       torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.4.attn_v.bias,         torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.4.attn_v.weight,       torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.5.attn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.5.ffn_down.weight,     torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.5.ffn_gate.weight,     torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.5.ffn_up.weight,       torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.5.ffn_norm.weight,     torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.5.attn_k.bias,         torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.5.attn_k.weight,       torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.5.attn_output.weight,  torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.5.attn_q.bias,         torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.5.attn_q.weight,       torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.5.attn_v.bias,         torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.5.attn_v.weight,       torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.6.attn_k.bias,         torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.6.attn_k.weight,       torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.6.attn_output.weight,  torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.6.attn_q.bias,         torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.6.attn_q.weight,       torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.6.attn_v.bias,         torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.6.attn_v.weight,       torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.10.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.10.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.10.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.10.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.10.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.10.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.10.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.10.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.10.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.10.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.10.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.10.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.11.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.11.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.11.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.11.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.11.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.11.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.11.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.11.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.11.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.11.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.11.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.11.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.12.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.12.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
```

```
INFO:hf-to-gguf:blk.12.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.12.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.12.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.12.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.12.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.12.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.12.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.12.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.12.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.12.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.13.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.13.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.13.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.13.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.13.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.13.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.13.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.13.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.13.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.13.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.13.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.13.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.14.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.14.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.14.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.14.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.14.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.14.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.14.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.14.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.14.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.14.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.14.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.14.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.15.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.15.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.15.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.15.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.15.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.15.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.15.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.6.attn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.6.ffn_down.weight,     torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.6.ffn_gate.weight,     torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.6.ffn_up.weight,       torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.6.ffn_norm.weight,     torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.7.attn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.7.ffn_down.weight,     torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.7.ffn_gate.weight,     torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.7.ffn_up.weight,       torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.7.ffn_norm.weight,     torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.7.attn_k.bias,         torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.7.attn_k.weight,       torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.7.attn_output.weight,  torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.7.attn_q.bias,         torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.7.attn_q.weight,       torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.7.attn_v.bias,         torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.7.attn_v.weight,       torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.8.attn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.8.ffn_down.weight,     torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.8.ffn_gate.weight,     torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.8.ffn_up.weight,       torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.8.ffn_norm.weight,     torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.8.attn_k.bias,         torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.8.attn_k.weight,       torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.8.attn_output.weight,  torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.8.attn_q.bias,         torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.8.attn_q.weight,       torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.8.attn_v.bias,         torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.8.attn_v.weight,       torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.9.attn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.9.ffn_down.weight,     torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.9.ffn_gate.weight,     torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.9.ffn_up.weight,       torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.9.ffn_norm.weight,     torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.9.attn_k.bias,         torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.9.attn_k.weight,       torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.9.attn_output.weight,  torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.9.attn_q.bias,         torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.9.attn_q.weight,       torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.9.attn_v.bias,         torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.9.attn_v.weight,       torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.15.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.15.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.15.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.15.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.15.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.16.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.16.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.16.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.16.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.16.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.16.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.16.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.16.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.16.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.16.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.16.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.16.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.17.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.17.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.17.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.17.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.17.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.17.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.17.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.17.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
```

```
INFO:hf-to-gguf:blk.17.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.17.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.17.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.17.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.18.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.18.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.18.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.18.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.18.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.18.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.18.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.18.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.18.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.18.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.18.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.18.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.19.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.19.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.19.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.19.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.19.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.19.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.19.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.19.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.19.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.19.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.19.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.19.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.20.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.20.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.20.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.20.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.20.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.20.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.20.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.20.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.20.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.20.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.20.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.20.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.21.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.21.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.21.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.21.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.21.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.21.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.21.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.21.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.21.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.21.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.21.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.21.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.22.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.22.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.22.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.22.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.22.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.22.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.22.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.22.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.22.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.22.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.22.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.22.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.23.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.23.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.23.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.23.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.23.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.23.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.23.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.23.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.23.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.23.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.23.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.23.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.24.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.24.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.24.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.24.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.24.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.24.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.24.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.24.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.24.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.24.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.24.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.24.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.25.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.25.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.25.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.25.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.25.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.25.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.25.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.25.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.25.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.25.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.25.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.25.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.26.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.26.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.26.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.26.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.26.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.26.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.26.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
```

```
INFO:hf-to-gguf:blk.26.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.26.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.26.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.26.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.26.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.27.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.27.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.27.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.27.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.27.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.27.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.27.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.27.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.27.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.27.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.27.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.27.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.28.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.28.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.28.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.28.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.28.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.28.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.28.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.28.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.28.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.28.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.28.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.28.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.29.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.29.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.29.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.29.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.29.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.29.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.29.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.29.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.29.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.29.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.29.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.29.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.30.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.30.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.30.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.30.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.30.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.30.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.30.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.30.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.30.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.30.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.30.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.30.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.31.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.31.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.31.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.31.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.31.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.31.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.31.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.31.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.31.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.31.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.31.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.31.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.32.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.32.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.32.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.32.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.32.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.32.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.32.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.32.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.32.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.32.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.32.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.32.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.33.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.33.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.33.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.33.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.33.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.33.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.33.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.33.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.33.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.33.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.33.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.33.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.34.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.34.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.34.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.34.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.34.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.34.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.34.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.34.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.34.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.34.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.34.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.34.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.35.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.35.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.35.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.35.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.35.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.35.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
```

```
INFO:hf-to-gguf:blk.35.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.35.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.35.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.35.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.35.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.35.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.36.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.36.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.36.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.36.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.36.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.36.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.36.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.36.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.36.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.36.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.36.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.36.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.37.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.37.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.37.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.37.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.37.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.37.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.37.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.37.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.37.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.37.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.37.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.37.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.38.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.38.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.38.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.38.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.38.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.38.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.38.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.38.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.38.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.38.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.38.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.38.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.39.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.39.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.39.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.39.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.39.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.39.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.39.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.39.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.39.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.39.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.39.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.39.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.40.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.40.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.40.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.40.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.40.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.40.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.40.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.40.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.40.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.40.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.40.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.40.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.41.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.41.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.41.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.41.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.41.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.41.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.41.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.41.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.41.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.41.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.41.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.41.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.42.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.42.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.42.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.42.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.42.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.42.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.42.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:output.weight,             torch.bfloat16 --> F16, shape = {5120, 152064}
INFO:hf-to-gguf:blk.42.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.42.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.42.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.42.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.42.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.43.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.43.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.43.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.43.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.43.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.43.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.43.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.43.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.43.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.43.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.43.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.43.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.44.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.44.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.44.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.44.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
```

```
INFO:hf-to-gguf:blk.44.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.44.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.44.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.44.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.44.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.44.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.44.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.44.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.45.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.45.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.45.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.45.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.45.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.45.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.45.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.45.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.45.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.45.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.45.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.45.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.46.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.46.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.46.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.46.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.46.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.46.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.46.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.46.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.46.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.46.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.46.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.46.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.47.attn_norm.weight,   torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.47.ffn_down.weight,    torch.bfloat16 --> F16, shape = {13824, 5120}
INFO:hf-to-gguf:blk.47.ffn_gate.weight,    torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.47.ffn_up.weight,      torch.bfloat16 --> F16, shape = {5120, 13824}
INFO:hf-to-gguf:blk.47.ffn_norm.weight,    torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.47.attn_k.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.47.attn_k.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:blk.47.attn_output.weight, torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.47.attn_q.bias,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:blk.47.attn_q.weight,      torch.bfloat16 --> F16, shape = {5120, 5120}
INFO:hf-to-gguf:blk.47.attn_v.bias,        torch.bfloat16 --> F32, shape = {1024}
INFO:hf-to-gguf:blk.47.attn_v.weight,      torch.bfloat16 --> F16, shape = {5120, 1024}
INFO:hf-to-gguf:output_norm.weight,        torch.bfloat16 --> F32, shape = {5120}
INFO:hf-to-gguf:Set meta model
INFO:hf-to-gguf:Set model parameters
INFO:hf-to-gguf:gguf: context length = 32768
INFO:hf-to-gguf:gguf: embedding length = 5120
INFO:hf-to-gguf:gguf: feed forward length = 13824
INFO:hf-to-gguf:gguf: head count = 40
INFO:hf-to-gguf:gguf: key-value head count = 8
INFO:hf-to-gguf:gguf: rope theta = 1000000.0
INFO:hf-to-gguf:gguf: rms norm epsilon = 1e-06
INFO:hf-to-gguf:gguf: file type = 1
INFO:hf-to-gguf:Set model quantization version
INFO:hf-to-gguf:Set model tokenizer
The tokenizer you are loading from '/workspace/code/MiPM_Senior_HF' with an incorrect regex pattern:
https://huggingface.co/mistralai/Mistral-Small-3.1-24B-Instruct-2503/discussions/84#69121093e8b480e709447d5e. This will
lead to incorrect tokenization. You should set the `fix_mistral_regex=True` flag when loading this tokenizer to fix this
issue.
INFO:gguf.vocab:Adding 151387 merge(s).
INFO:gguf.vocab:Setting special token type eos to 151645
INFO:gguf.vocab:Setting special token type pad to 151665
INFO:gguf.vocab:Setting chat_template to {% for message in messages %}{% if message['role'] == 'user' %}
{{'<|im_start|>user
' + message['content'] + '<|im_end|>
'}}{% elif message['role'] == 'assistant' %}{{'<|im_start|>assistant
' + message['content'] + '<|im_end|>
' }}{% else %}{{ '<|im_start|>system
' + message['content'] + '<|im_end|>
' }}{% endif %}{% endfor %}{% if add_generation_prompt %}{{ '<|im_start|>assistant
' }}{% endif %}
INFO:gguf.gguf_writer:Writing the following files:
INFO:gguf.gguf_writer:/workspace/model_temp_f16.gguf: n_tensors = 579, total_size = 29.5G
Writing: 100%|██████████| 29.5G/29.5G [02:53<00:00, 170Mbyte/s]
INFO:hf-to-gguf:Model successfully exported to /workspace/model_temp_f16.gguf

✅ Conversión a F16 completada.
🛠️ PASO 2: Compilando llama-quantize con CMake...
-- The C compiler identification is GNU 11.4.0
-- The CXX compiler identification is GNU 11.4.0
-- Detecting C compiler ABI info
-- Detecting C compiler ABI info - done
-- Check for working C compiler: /usr/bin/cc - skipped
-- Detecting C compile features
-- Detecting C compile features - done
-- Detecting CXX compiler ABI info
-- Detecting CXX compiler ABI info - done
-- Check for working CXX compiler: /usr/bin/c++ - skipped
-- Detecting CXX compile features
-- Detecting CXX compile features - done
-- Found Git: /usr/bin/git (found version "2.34.1")

  [0mCMAKE_BUILD_TYPE=Release [0m

-- The ASM compiler identification is GNU
-- Found assembler: /usr/bin/cc
-- Looking for pthread.h
-- Looking for pthread.h - found
-- Performing Test CMAKE_HAVE_LIBC_PTHREAD
-- Performing Test CMAKE_HAVE_LIBC_PTHREAD - Success
-- Found Threads: TRUE
-- Warning: ccache not found - consider installing it for faster compilation or disable this warning with GGML_CCACHE=OFF
-- CMAKE_SYSTEM_PROCESSOR: x86_64
```

```
-- GGML_SYSTEM_ARCH: x86
-- Including CPU backend
-- Found OpenMP_C: -fopenmp (found version "4.5")
-- Found OpenMP_CXX: -fopenmp (found version "4.5")
-- Found OpenMP: TRUE (found version "4.5")
-- x86 detected
-- Adding CPU backend variant ggml-cpu: -march=native
-- Found CUDAToolkit: /usr/local/cuda/include (found version "12.4.131")
-- CUDA Toolkit found
-- The CUDA compiler identification is NVIDIA 12.4.131
-- Detecting CUDA compiler ABI info
-- Detecting CUDA compiler ABI info - done
-- Check for working CUDA compiler: /usr/local/cuda/bin/nvcc - skipped
-- Detecting CUDA compile features
-- Detecting CUDA compile features - done
-- Using CMAKE_CUDA_ARCHITECTURES=50-virtual;61-virtual;70-virtual;75-virtual;80-virtual;86-real;89-real
CMAKE_CUDA_ARCHITECTURES_NATIVE=
-- CUDA host compiler is GNU 11.4.0
-- Including CUDA backend
-- ggml version: 0.9.7
-- ggml commit:  10b26ee23
-- Could NOT find OpenSSL, try to set the path to OpenSSL root folder in the system variable OPENSSL_ROOT_DIR (missing:
OPENSSL_CRYPTO_LIBRARY OPENSSL_INCLUDE_DIR)


  [33mCMake Warning at vendor/cpp-httplib/CMakeLists.txt:150 (message):
   OpenSSL not found, HTTPS support disabled

  [0m


-- Generating embedded license file for target: common
-- Configuring done
-- Generating done
-- Build files have been written to: /workspace/code/llama.cpp/build
[  0%]  [32mBuilding CXX object common/CMakeFiles/build_info.dir/build-info.cpp.o [0m
[  0%]  [32mBuilding CXX object vendor/cpp-httplib/CMakeFiles/cpp-httplib.dir/httplib.cpp.o [0m
[  1%]  [32mBuilding CXX object ggml/src/CMakeFiles/ggml-base.dir/ggml.cpp.o [0m
[  1%]  [32mBuilding C object ggml/src/CMakeFiles/ggml-base.dir/ggml.c.o [0m
[  1%]  [32mBuilding C object ggml/src/CMakeFiles/ggml-base.dir/ggml-alloc.c.o [0m
[  1%]  [32mBuilding CXX object ggml/src/CMakeFiles/ggml-base.dir/ggml-opt.cpp.o [0m
[  1%]  [32mBuilding CXX object ggml/src/CMakeFiles/ggml-base.dir/ggml-threading.cpp.o [0m
[  1%]  [32mBuilding CXX object ggml/src/CMakeFiles/ggml-base.dir/ggml-backend.cpp.o [0m
[  1%]  [32mBuilding CXX object ggml/src/CMakeFiles/ggml-base.dir/gguf.cpp.o [0m
[  1%]  [32mBuilding C object ggml/src/CMakeFiles/ggml-base.dir/ggml-quants.c.o [0m
[  1%] Built target build_info
[  3%]  [32m [1mLinking CXX shared library ../../bin/libggml-base.so [0m
[  3%] Built target ggml-base
[  3%]  [32mBuilding C object ggml/src/CMakeFiles/ggml-cpu.dir/ggml-cpu/ggml-cpu.c.o [0m
[  3%]  [32mBuilding CXX object ggml/src/CMakeFiles/ggml-cpu.dir/ggml-cpu/ggml-cpu.cpp.o [0m
[  3%]  [32mBuilding CXX object ggml/src/CMakeFiles/ggml-cpu.dir/ggml-cpu/hbm.cpp.o [0m
[  3%]  [32mBuilding CXX object ggml/src/CMakeFiles/ggml-cpu.dir/ggml-cpu/repack.cpp.o [0m
[  3%]  [32mBuilding C object ggml/src/CMakeFiles/ggml-cpu.dir/ggml-cpu/quants.c.o [0m
[  5%]  [32mBuilding CXX object ggml/src/CMakeFiles/ggml-cpu.dir/ggml-cpu/amx/amx.cpp.o [0m
[  5%]  [32mBuilding CXX object ggml/src/CMakeFiles/ggml-cpu.dir/ggml-cpu/amx/mmq.cpp.o [0m
[  5%]  [32mBuilding CXX object ggml/src/CMakeFiles/ggml-cpu.dir/ggml-cpu/binary-ops.cpp.o [0m
[  5%]  [32mBuilding CXX object ggml/src/CMakeFiles/ggml-cpu.dir/ggml-cpu/traits.cpp.o [0m
[  6%]  [32mBuilding CXX object ggml/src/CMakeFiles/ggml-cpu.dir/ggml-cpu/unary-ops.cpp.o [0m
[  6%]  [32mBuilding CXX object ggml/src/CMakeFiles/ggml-cpu.dir/ggml-cpu/ops.cpp.o [0m
[  6%]  [32mBuilding CXX object ggml/src/CMakeFiles/ggml-cpu.dir/ggml-cpu/llamafile/sgemm.cpp.o [0m
[  6%]  [32mBuilding CXX object ggml/src/CMakeFiles/ggml-cpu.dir/ggml-cpu/vec.cpp.o [0m
[  6%]  [32mBuilding CXX object ggml/src/CMakeFiles/ggml-cpu.dir/ggml-cpu/arch/x86/repack.cpp.o [0m
[  6%]  [32mBuilding C object ggml/src/CMakeFiles/ggml-cpu.dir/ggml-cpu/arch/x86/quants.c.o [0m
[  8%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/arange.cu.o [0m
[  8%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/add-id.cu.o [0m
[  8%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/argmax.cu.o [0m
[  8%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/acc.cu.o [0m
[  8%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/argsort.cu.o [0m
[ 10%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/binbcast.cu.o [0m
[ 10%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/concat.cu.o [0m
[ 10%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/conv-transpose-1d.cu.o [0m
[ 12%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/conv2d-transpose.cu.o [0m
[ 12%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/clamp.cu.o [0m
[ 12%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/conv2d-dw.cu.o [0m
[ 12%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/convert.cu.o [0m
[ 12%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/conv2d.cu.o [0m
[ 12%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/count-equal.cu.o [0m
[ 12%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/cpy.cu.o [0m
[ 12%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/cross-entropy-loss.cu.o [0m
[ 12%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/cumsum.cu.o [0m
[ 13%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/fattn-tile.cu.o [0m
[ 13%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/fattn-wmma-f16.cu.o [0m
[ 13%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/diagmask.cu.o [0m
[ 13%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/fattn.cu.o [0m
[ 13%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/diag.cu.o [0m
[ 15%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/getrows.cu.o [0m
[ 15%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/fill.cu.o [0m
[ 15%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/ggml-cuda.cu.o [0m
[ 15%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/gla.cu.o [0m
[ 15%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/im2col.cu.o [0m
[ 15%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/mean.cu.o [0m
[ 17%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/mmq.cu.o [0m
[ 17%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/mmid.cu.o [0m
[ 17%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/mmvf.cu.o [0m
[ 17%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/mmf.cu.o [0m
[ 17%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/mmvq.cu.o [0m
[ 17%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/norm.cu.o [0m
[ 18%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/opt-step-sgd.cu.o [0m
[ 18%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/out-prod.cu.o [0m
[ 18%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/pad.cu.o [0m
[ 18%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/opt-step-adamw.cu.o [0m
[ 20%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/pad_reflect_1d.cu.o [0m
[ 20%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/quantize.cu.o [0m
[ 20%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/roll.cu.o [0m
[ 20%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/rope.cu.o [0m
[ 20%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/scale.cu.o [0m
[ 20%]  [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/pool2d.cu.o [0m
```

```
[ 20%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/set-rows.cu.o [0m
[ 22%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/softcap.cu.o [0m
[ 22%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/solve_tri.cu.o [0m
[ 22%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/softmax.cu.o [0m
[ 22%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/set.cu.o [0m
[ 22%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/ssm-conv.cu.o [0m
[ 24%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/sum.cu.o [0m
[ 24%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/sumrows.cu.o [0m
[ 24%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/ssm-scan.cu.o [0m
[ 24%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/top-k.cu.o [0m
[ 24%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/topk-moe.cu.o [0m
[ 24%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/tri.cu.o [0m
[ 25%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/unary.cu.o [0m
[ 25%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/upscale.cu.o [0m
[ 25%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/fattn-tile-instance-dkq112-
dv112.cu.o [0m
[ 25%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/wkv.cu.o [0m
[ 25%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/fattn-tile-instance-dkq128-
dv128.cu.o [0m
[ 27%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/tsembd.cu.o [0m
[ 27%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/fattn-tile-instance-dkq256-
dv256.cu.o [0m
[ 27%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/fattn-tile-instance-dkq40-
dv40.cu.o [0m
[ 27%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/fattn-tile-instance-dkq576-
dv512.cu.o [0m
[ 27%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/fattn-tile-instance-dkq64-
dv64.cu.o [0m
[ 27%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/fattn-tile-instance-dkq72-
dv72.cu.o [0m
[ 29%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/fattn-mma-f16-instance-
ncols1_1-ncols2_16.cu.o [0m
[ 29%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/fattn-tile-instance-dkq96-
dv96.cu.o [0m
[ 29%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/fattn-mma-f16-instance-
ncols1_1-ncols2_32.cu.o [0m
[ 29%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/fattn-tile-instance-dkq80-
dv80.cu.o [0m
[ 29%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/fattn-mma-f16-instance-
ncols1_1-ncols2_8.cu.o [0m
[ 31%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/fattn-mma-f16-instance-
ncols1_16-ncols2_1.cu.o [0m
[ 31%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/fattn-mma-f16-instance-
ncols1_16-ncols2_4.cu.o [0m
[ 31%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/fattn-mma-f16-instance-
ncols1_2-ncols2_16.cu.o [0m
[ 31%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/fattn-mma-f16-instance-
ncols1_16-ncols2_2.cu.o [0m
[ 31%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/fattn-mma-f16-instance-
ncols1_2-ncols2_32.cu.o [0m
[ 32%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/fattn-mma-f16-instance-
ncols1_2-ncols2_4.cu.o [0m
[ 32%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/fattn-mma-f16-instance-
ncols1_32-ncols2_1.cu.o [0m
[ 32%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/fattn-mma-f16-instance-
ncols1_32-ncols2_2.cu.o [0m
[ 32%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/fattn-mma-f16-instance-
ncols1_2-ncols2_8.cu.o [0m
[ 32%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/fattn-mma-f16-instance-
ncols1_4-ncols2_16.cu.o [0m
[ 32%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/fattn-mma-f16-instance-
ncols1_4-ncols2_2.cu.o [0m
[ 32%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/fattn-mma-f16-instance-
ncols1_4-ncols2_4.cu.o [0m
[ 34%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/fattn-mma-f16-instance-
ncols1_64-ncols2_1.cu.o [0m
[ 34%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/fattn-mma-f16-instance-
ncols1_4-ncols2_8.cu.o [0m
[ 34%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/fattn-mma-f16-instance-
ncols1_8-ncols2_1.cu.o [0m
[ 34%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/fattn-mma-f16-instance-
ncols1_8-ncols2_2.cu.o [0m
[ 34%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/fattn-mma-f16-instance-
ncols1_8-ncols2_4.cu.o [0m
[ 34%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/fattn-mma-f16-instance-
ncols1_8-ncols2_8.cu.o [0m
[ 36%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/mmq-instance-iq2_s.cu.o [0m
[ 36%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/mmq-instance-iq3_s.cu.o [0m
[ 36%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/mmq-instance-iq2_xxs.cu.o [0m
[ 36%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/mmq-instance-iq2_xs.cu.o [0m
[ 37%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/mmq-instance-iq1_s.cu.o [0m
[ 37%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/mmq-instance-iq3_xxs.cu.o [0m
[ 37%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/mmq-instance-iq4_nl.cu.o [0m
[ 37%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/mmq-instance-iq4_xs.cu.o [0m
[ 37%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/mmq-instance-mxfp4.cu.o [0m
[ 37%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/mmq-instance-q2_k.cu.o [0m
[ 37%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/mmq-instance-q3_k.cu.o [0m
[ 39%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/mmq-instance-q4_k.cu.o [0m
[ 39%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/mmq-instance-q4_1.cu.o [0m
[ 39%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/mmq-instance-q5_0.cu.o [0m
[ 39%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/mmq-instance-q4_0.cu.o [0m
[ 39%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/mmq-instance-q5_1.cu.o [0m
[ 41%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/mmq-instance-q6_k.cu.o [0m
[ 41%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/mmq-instance-q8_0.cu.o [0m
[ 41%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/mmq-instance-q5_k.cu.o [0m
[ 41%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/mmf-instance-ncols_1.cu.o [0m
[ 41%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/mmf-instance-ncols_10.cu.o [0m
[ 41%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/mmf-instance-ncols_11.cu.o [0m
[ 43%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/mmf-instance-ncols_13.cu.o [0m
[ 43%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/mmf-instance-ncols_14.cu.o [0m
[ 43%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/mmf-instance-ncols_15.cu.o [0m
[ 43%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/mmf-instance-ncols_12.cu.o [0m
[ 43%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/mmf-instance-ncols_2.cu.o [0m
[ 43%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/mmf-instance-ncols_16.cu.o [0m
[ 44%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/mmf-instance-ncols_4.cu.o [0m
[ 44%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/mmf-instance-ncols_3.cu.o [0m
[ 46%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/mmf-instance-ncols_6.cu.o [0m
```

```
[ 46%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/mmf-instance-ncols_7.cu.o [0m
[ 46%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/mmf-instance-ncols_5.cu.o [0m
[ 46%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/mmf-instance-ncols_9.cu.o [0m
[ 46%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/mmf-instance-ncols_8.cu.o [0m
[ 46%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/fattn-vec-instance-q4_0-
q4_0.cu.o [0m
[ 46%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/fattn-vec-instance-q8_0-
q8_0.cu.o [0m
[ 46%] [32mBuilding CUDA object ggml/src/ggml-cuda/CMakeFiles/ggml-cuda.dir/template-instances/fattn-vec-instance-f16-
f16.cu.o [0m
[ 46%] [32m [1mLinking CXX static library libcpp-httplib.a [0m
[ 46%] Built target cpp-httplib
[ 46%] [32m [1mLinking CXX shared library ../../bin/libggml-cpu.so [0m
[ 46%] Built target ggml-cpu
```

In [ ]:

Exported with runcell — convert notebooks to HTML or PDF anytime at runcell.dev.