## NAME

cmfinetune − train LoRA adapters on a frozen model and merge them

## SYNOPSIS

**cmfinetune** [*OPTION*]... *MODEL DATA.jsonl*
**cmfinetune** [*OPTION*]... *MODEL* **−−merge** *ADAPTER*

## DESCRIPTION

**cmfinetune** trains low-rank adapters (LoRA) on the projections of a frozen model and writes them as safetensors. The base weights keep their file encoding; the forward pass uses the inference kernels. With **−−merge**, it writes the model with an adapter added as a GGUF instead.

**cmgenerate**(1) and **cmagent**(1) load an adapter with **−−lora**, which adds it to the weights in memory before any offload.

### Supported models

Every architecture the library runs: attention, latent attention with or without the sparse indexer, gated-delta layers (Qwen3-Next, Qwen3.5, Kimi), hyper-connections, and dense, xIELU or routed feed-forward layers. Frozen parts propagate the gradient; adapters go on the projections. In a mixture-of-experts layer the routed experts stay frozen and **gate,up,down** adapt the shared expert. In a gated-delta layer **q** adapts the fused **attn_qkv**; in a latent-attention layer **q,k,v** adapt **attn_q_b**, **attn_kv_a** and **attn_kv_b**.

## DATA FORMAT

One JSON object per line:

**{"tokens": [...]}**
: Token ids; every token is trained. For models without a vocabulary.

**{"text": "..."}**
: Every token is trained.

**{"prompt": "...", "completion": "..."}**
: Only the completion is trained.

**{"messages": [{"role": "...", "content": "..."}, ...]}**
: Rendered with the model’s chat template; assistant turns and their end-of-turn token are trained.

Examples longer than **−−seq−len** tokens are truncated.

## OPTIONS

### Training

**−o**, **−−out** *PATH*
: Adapter output. Default: *MODEL*.lora.safetensors.

**−r**, **−−rank** *N*
: Adapter rank. Default: 8.

**−−alpha** *F*
: The adapter is scaled by alpha/rank. Default: 2*rank.

**−−targets** *LIST*
: Comma-separated subset of q,k,v,o,gate,up,down, or attention, mlp, all. Default: all.

**−−lr** *F*
: Peak learning rate. Default: 1e-4.

**−−epochs** *N*
: Passes over the data. Default: 1.

**−−steps** *N*
: Stop after *N* optimiser steps.

**−−accumulate** *N*
: Examples per optimiser step (gradient accumulation). Default: 1.

**−−seq−len** *N*
: Truncate examples to *N* tokens. Default: 1024.

**−−warmup** *N*
: Linear warmup steps, then cosine decay to lr/10. Default: 10.

**−−weight−decay** *F*
: AdamW weight decay. Default: 0.

**−−clip** *F*
: Global gradient-norm clip; 0 disables. Default: 1.

**−s**, **−−seed** *N*
: Seed for shuffling and adapter initialisation. Default: 1.

**−−eval** *PATH*
: Report the loss on *PATH* (same format as the data) after every epoch.

**−−save−every** *N*
: Also write the adapter every *N* steps.

**−−resume** *ADAPTER*
: Start from *ADAPTER*; the optimiser state starts fresh.

**−−keep** *GIB*
: Memory for layer activations kept from the forward pass; layers beyond it are recomputed in the backward pass. Default: 4.

**Placement**

**−−gpu**
: Run the frozen matrix products, forward and backward, on the GPU; the model and the transposed copies are uploaded.

**−−gpu−backend** *API*
: auto, cuda, metal or vulkan. Implies **−−gpu**.

**−−gpu−device** *N[,N...]*
: Vulkan physical device indices, or CUDA ordinals with **−−gpu−backend cuda**. With several, the layers are split across the devices in proportion to their free memory.

**−−gpu−split** *R[,R...]*
: Each device's share of the layers instead, like 3,1. A device still takes no more than fits.

**−−gpu−as** *FMT*
: Quantise the model to *FMT* for the GPU.

**−−backward−as** *FMT*
: Format of the transposed weight copies the backward pass multiplies by: Q8_0, BF16, F32, or **none** to decode the weights on the fly and hold no copies (about 12% slower). Default: Q8_0, about 1.6 times the size of a Q4 model.

**Merging**

**−−merge** *ADAPTER*
: Write *MODEL* with *ADAPTER* merged and exit.

**−o**, **−−out** *PATH*
: Output. Default: *MODEL*.merged.gguf.

**−−as** *FMT*
: Format of the float matrices the adapter does not change: Q4_K, Q5_K, Q6_K or Q8_0. Default: Q8_0.

**−−adapted−as** *FMT*
: Format of the matrices the adapter changes. Default: Q8_0; narrower formats lose much of a low-rank update.

## ENVIRONMENT

**CMGENERATE_THREADS**
: Worker threads; **−−threads** *N* overrides it.

**CMGENERATE_PANEL_TOKENS**
: As in **cmgenerate**(1): batches of at least this many tokens use panel products (default 24; 0 never). Adapter products of a million multiply-adds or more always do.

## EXAMPLES

cmfinetune model.gguf train.jsonl −−eval eval.jsonl −−epochs 3 −−lr 1e−3
cmgenerate model.gguf −−lora model.lora.safetensors "Hello"
cmfinetune model.gguf −−merge model.lora.safetensors −o tuned.gguf

## SEE ALSO

**cmgenerate**(1), **cmagent**(1)