```
Last login: Mon Jul 21 13:43:13 on ttys008
/Users/andrebaker/Desktop/
chat_llama.command ; exit;
andrebaker@Andres-MacBook-Pro ~ % /
Users/andrebaker/Desktop/
chat_llama.command ; exit;
🧠 LLaMA server started (PID: 14586)
🎤 Launching Jarvis voice chat...
build: 5948 (b4efd77f) with Apple clang
version 17.0.0 (clang-1700.0.13.5) for
arm64-apple-darwin24.5.0
system info: n_threads = 6, n_threads_batch
= 6, total_threads = 10

system_info: n_threads = 6
(n_threads_batch = 6) / 10 | Metal :
EMBED_LIBRARY = 1 | CPU : NEON = 1 |
ARM_FMA = 1 | FP16_VA = 1 | DOTPROD = 1 |
LLAMAFILE = 1 | ACCELERATE = 1 | REPACK
= 1 |

main: binding port with default address family
main: HTTP server is listening, hostname:
127.0.0.1, port: 8080, http threads: 9
main: loading model
```

```
srv    load_model: loading model 'models/
meta-llama-3-8b-instruct.Q4_K_M.gguf'
llama_model_load_from_file_impl: using
device Metal (Apple M1 Pro) - 10922 MiB
free
llama_model_loader: loaded meta data with
21 key-value pairs and 291 tensors from
models/meta-llama-3-8b-
instruct.Q4_K_M.gguf (version GGUF V3
(latest))
llama_model_loader: Dumping metadata
keys/values. Note: KV overrides do not apply
in this output.
llama_model_loader: - kv   0:
general.architecture str            = llama
llama_model_loader: - kv   1:
general.name str            = Meta-Llama-3-8B-
Instruct
llama_model_loader: - kv   2:
llama.block_count u32            = 32
llama_model_loader: - kv   3:
llama.context_length u32            = 8192
llama_model_loader: - kv   4:
llama.embedding_length u32            = 4096
llama_model_loader: - kv   5:
```

```
llama.feed_forward_length u32              =
14336
llama_model_loader: - kv   6:
llama.attention.head_count u32             = 32
llama_model_loader: - kv   7:
llama.attention.head_count_kv u32              =
8
llama_model_loader: - kv   8:
llama.rope.freq_base f32             =
500000.000000
llama_model_loader: - kv   9:
llama.attention.layer_norm_rms_epsilon f32
= 0.000010
llama_model_loader: - kv  10:
general.file_type u32             = 15
llama_model_loader: - kv  11:
llama.vocab_size u32             = 128256
llama_model_loader: - kv  12:
llama.rope.dimension_count u32             =
128
llama_model_loader: - kv  13:
tokenizer.ggml.model str             = gpt2
llama_model_loader: - kv  14:
tokenizer.ggml.tokens arr[str,128256]  = ["!",
"\"", "#", "$", "%", "&", "'", ...
```

```
llama_model_loader: - kv  15:
tokenizer.ggml.token_type arr[i32,128256]  =
[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
llama_model_loader: - kv  16:
tokenizer.ggml.merges arr[str,280147]  = ["Ġ
Ġ", "Ġ ĠĠĠ", "ĠĠ ĠĠ", "...
llama_model_loader: - kv  17:
tokenizer.ggml.bos_token_id u32            =
128000
llama_model_loader: - kv  18:
tokenizer.ggml.eos_token_id u32            =
128001
llama_model_loader: - kv  19:
tokenizer.chat_template str            = {% set
loop_messages = messages %}{% ×××
llama_model_loader: - kv  20:
general.quantization_version u32            = 2
llama_model_loader: - type  f32:   65 tensors
llama_model_loader: - type q4_K:  193
tensors
llama_model_loader: - type q6_K:   33
tensors
print_info: file format = GGUF V3 (latest)
print_info: file type   = Q4_K - Medium
print_info: file size   = 4.58 GiB (4.89 BPW)
```

```
load: missing pre-tokenizer type, using: 'default'
load:
load: ***************************************
load: GENERATION QUALITY WILL BE DEGRADED!
load: CONSIDER REGENERATING THE MODEL
load: ***************************************
load:
load: special tokens cache size = 256
load: token to piece cache size = 0.8000 MB
print_info: arch             = llama
print_info: vocab_only       = 0
print_info: n_ctx_train      = 8192
print_info: n_embd           = 4096
print_info: n_layer          = 32
print_info: n_head           = 32
print_info: n_head_kv        = 8
print_info: n_rot            = 128
print_info: n_swa            = 0
print_info: is_swa_any       = 0
print_info: n_embd_head_k    = 128
print_info: n_embd_head_v    = 128
print_info: n_gqa            = 4
```

```
print_info: n_embd_k_gqa     = 1024
print_info: n_embd_v_gqa     = 1024
print_info: f_norm_eps       = 0.0e+00
print_info: f_norm_rms_eps   = 1.0e-05
print_info: f_clamp_kqv      = 0.0e+00
print_info: f_max_alibi_bias = 0.0e+00
print_info: f_logit_scale    = 0.0e+00
print_info: f_attn_scale     = 0.0e+00
print_info: n_ff             = 14336
print_info: n_expert         = 0
print_info: n_expert_used    = 0
print_info: causal attn      = 1
print_info: pooling type     = 0
print_info: rope type        = 0
print_info: rope scaling     = linear
print_info: freq_base_train  = 500000.0
print_info: freq_scale_train = 1
print_info: n_ctx_orig_yarn  = 8192
print_info: rope_finetuned   = unknown
print_info: model type       = 8B
print_info: model params     = 8.03 B
print_info: general.name     = Meta-
Llama-3-8B-Instruct
print_info: vocab type       = BPE
print_info: n_vocab          = 128256
```

```
print_info: n_merges         = 280147
print_info: BOS token        = 128000 '<|
begin_of_text|>'
print_info: EOS token        = 128001 '<|
end_of_text|>'
print_info: EOT token        = 128009 '<|eot_id|
>'
print_info: LF token         = 198 'Ċ'
print_info: EOG token        = 128001 '<|
end_of_text|>'
print_info: EOG token        = 128009 '<|
eot_id|>'
print_info: max token length = 256
load_tensors: loading model tensors, this can
take a while... (mmap = true)
load_tensors: offloading 32 repeating layers
to GPU
load_tensors: offloading output layer to GPU
load_tensors: offloaded 33/33 layers to GPU
load_tensors: Metal_Mapped model buffer
size =  4403.50 MiB
load_tensors:   CPU_Mapped model buffer
size =   281.81 MiB
.........................................................................
....
```

```
llama_context: constructing llama_context
llama_context: non-unified KV cache requires ggml_set_rows() - forcing unified KV cache
llama_context: n_seq_max     = 1
llama_context: n_ctx         = 4096
llama_context: n_ctx_per_seq = 4096
llama_context: n_batch       = 2048
llama_context: n_ubatch      = 512
llama_context: causal_attn   = 1
llama_context: flash_attn    = 0
llama_context: kv_unified    = true
llama_context: freq_base     = 500000.0
llama_context: freq_scale    = 1
llama_context: n_ctx_per_seq (4096) < n_ctx_train (8192) -- the full capacity of the model will not be utilized
ggml_metal_init: allocating
ggml_metal_init: found device: Apple M1 Pro
ggml_metal_init: picking default device: Apple M1 Pro
ggml_metal_load_library: using embedded metal library
ggml_metal_init: GPU name:   Apple M1 Pro
ggml_metal_init: GPU family: MTLGPUFamilyApple7  (1007)
```

```
ggml_metal_init: GPU family:
MTLGPUFamilyCommon3 (3003)
ggml_metal_init: GPU family:
MTLGPUFamilyMetal3  (5001)
ggml_metal_init: simdgroup reduction   = true
ggml_metal_init: simdgroup matrix mul. =
true
ggml_metal_init: has residency sets    = true
ggml_metal_init: has bfloat            = true
ggml_metal_init: use bfloat            = false
ggml_metal_init: hasUnifiedMemory      = true
ggml_metal_init:
recommendedMaxWorkingSetSize  =
11453.25 MB
ggml_metal_init: skipping
kernel_get_rows_bf16                    (not
supported)
ggml_metal_init: skipping
kernel_set_rows_bf16                    (not
supported)
ggml_metal_init: skipping
kernel_mul_mv_bf16_f32                  (not
supported)
ggml_metal_init: skipping
kernel_mul_mv_bf16_f32_c4               (not
supported)
ggml_metal_init: skipping
kernel_mul_mv_bf16_f32_1row          (not
supported)
ggml_metal_init: skipping
kernel_mul_mv_bf16_f32_l4            (not
supported)
ggml_metal_init: skipping
kernel_mul_mv_bf16_bf16              (not
supported)
ggml_metal_init: skipping
kernel_mul_mv_id_bf16_f32            (not
supported)
ggml_metal_init: skipping
kernel_mul_mm_bf16_f32               (not
supported)
ggml_metal_init: skipping
kernel_mul_mm_id_bf16_f16            (not
supported)
ggml_metal_init: skipping
kernel_flash_attn_ext_bf16_h64       (not
supported)
ggml_metal_init: skipping
kernel_flash_attn_ext_bf16_h80       (not
supported)
```

```
ggml_metal_init: skipping
kernel_flash_attn_ext_bf16_h96          (not
supported)
ggml_metal_init: skipping
kernel_flash_attn_ext_bf16_h112         (not
supported)
ggml_metal_init: skipping
kernel_flash_attn_ext_bf16_h128         (not
supported)
ggml_metal_init: skipping
kernel_flash_attn_ext_bf16_h192         (not
supported)
ggml_metal_init: skipping
kernel_flash_attn_ext_bf16_hk192_hv128
(not supported)
ggml_metal_init: skipping
kernel_flash_attn_ext_bf16_h256         (not
supported)
ggml_metal_init: skipping
kernel_flash_attn_ext_bf16_hk576_hv512
(not supported)
ggml_metal_init: skipping
kernel_flash_attn_ext_vec_bf16_h64      (not
supported)
ggml_metal_init: skipping
```

```
kernel_flash_attn_ext_vec_bf16_h96       (not
supported)
ggml_metal_init: skipping
kernel_flash_attn_ext_vec_bf16_h128
(not supported)
ggml_metal_init: skipping
kernel_flash_attn_ext_vec_bf16_h192
(not supported)
ggml_metal_init: skipping
kernel_flash_attn_ext_vec_bf16_hk192_hv1
28 (not supported)
ggml_metal_init: skipping
kernel_flash_attn_ext_vec_bf16_h256
(not supported)
ggml_metal_init: skipping
kernel_flash_attn_ext_vec_bf16_hk576_hv5
12 (not supported)
ggml_metal_init: skipping
kernel_cpy_f32_bf16                  (not
supported)
ggml_metal_init: skipping
kernel_cpy_bf16_f32                  (not
supported)
ggml_metal_init: skipping
kernel_cpy_bf16_bf16                  (not
supported)
llama_context:       CPU  output buffer size =
0.49 MiB
llama_kv_cache_unified:      Metal KV buffer
size =   512.00 MiB
llama_kv_cache_unified: size =  512.00 MiB
(  4096 cells,  32 layers,  1/ 1 seqs), K (f16):
256.00 MiB, V (f16):  256.00 MiB
llama_kv_cache_unified:
LLAMA_SET_ROWS=0, using old ggml_cpy()
method for backwards compatibility
llama_context:      Metal compute buffer size
=   296.00 MiB
llama_context:       CPU compute buffer size
=    16.01 MiB
llama_context: graph nodes  = 1190
llama_context: graph splits = 2
common_init_from_params: added <|
end_of_text|> logit bias = -inf
common_init_from_params: added <|eot_id|
> logit bias = -inf
common_init_from_params: setting
dry_penalty_last_n to ctx_size = 4096
common_init_from_params: warming up the
model with an empty run - please wait ... (--
no-warmup to disable)
srv         init: initializing slots, n_slots = 1
slot        init: id  0 | task -1 | new slot
n_ctx_slot = 4096
main: model loaded
main: chat template, chat_template: {% set
loop_messages = messages %}{% for
message in loop_messages %}{% set content
= '<|start_header_id|>' + message['role'] +
'<|end_header_id|>

'+ message['content'] | trim + '<|eot_id|>' %}
{% if loop×index0 == 0 %}{% set content =
bos_token + content %}{% endif %}
{{ content }}{% endfor %}{% if
add_generation_prompt %}{{ '<|
start_header_id|>assistant<|end_header_id|
>

' }}{% endif %}, example_format: '<|
start_header_id|>system<|end_header_id|>

You are a helpful assistant<|eot_id|><|
start_header_id|>user<|end_header_id|>
```

```
Hello<|eot_id|><|start_header_id|
>assistant<|end_header_id|>

Hi there<|eot_id|><|start_header_id|>user<|
end_header_id|>

How are you?<|eot_id|><|start_header_id|
>assistant<|end_header_id|>

'
main: server is listening on http://
127.0.0.1:8080 - starting the main loop
srv  update_slots: all slots are idle
llama_model_load_from_file_impl: using
device Metal (Apple M1 Pro) - 10922 MiB
free
llama_model_loader: loaded meta data with
21 key-value pairs and 291 tensors from /
Users/andrebaker/llama.cpp/models/meta-
llama-3-8b-instruct.Q4_K_M.gguf (version
GGUF V3 (latest))
llama_model_loader: Dumping metadata
keys/values. Note: KV overrides do not apply
in this output.
```

```
llama_model_loader: - kv   0:
general.architecture str              = llama
llama_model_loader: - kv   1:
general.name str              = Meta-Llama-3-8B-Instruct
llama_model_loader: - kv   2:
llama.block_count u32              = 32
llama_model_loader: - kv   3:
llama.context_length u32              = 8192
llama_model_loader: - kv   4:
llama.embedding_length u32              = 4096
llama_model_loader: - kv   5:
llama.feed_forward_length u32              =
14336
llama_model_loader: - kv   6:
llama.attention.head_count u32              = 32
llama_model_loader: - kv   7:
llama.attention.head_count_kv u32              =
8
llama_model_loader: - kv   8:
llama.rope.freq_base f32              =
500000.000000
llama_model_loader: - kv   9:
llama.attention.layer_norm_rms_epsilon f32
= 0.000010
```

```
llama_model_loader: - kv  10:
general.file_type u32              = 15
llama_model_loader: - kv  11:
llama.vocab_size u32               = 128256
llama_model_loader: - kv  12:
llama.rope.dimension_count u32               =
128
llama_model_loader: - kv  13:
tokenizer.ggml.model str              = gpt2
llama_model_loader: - kv  14:
tokenizer.ggml.tokens arr[str,128256]  = ["!",
"\"", "#", "$", "%", "&", "'", ...
llama_model_loader: - kv  15:
tokenizer.ggml.token_type arr[i32,128256]  =
[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
llama_model_loader: - kv  16:
tokenizer.ggml.merges arr[str,280147]  = ["Ġ
Ġ", "Ġ ĠĠĠ", "ĠĠ ĠĠ", "...
llama_model_loader: - kv  17:
tokenizer.ggml.bos_token_id u32              =
128000
llama_model_loader: - kv  18:
tokenizer.ggml.eos_token_id u32              =
128001
llama_model_loader: - kv  19:
```

```
tokenizer.chat_template str             = {% set
loop_messages = messages %}{% ×××
llama_model_loader: - kv  20:
general.quantization_version u32             = 2
llama_model_loader: - type  f32:   65 tensors
llama_model_loader: - type q4_K:  193
tensors
llama_model_loader: - type q6_K:   33
tensors
print_info: file format = GGUF V3 (latest)
print_info: file type   = Q4_K - Medium
print_info: file size   = 4.58 GiB (4.89 BPW)
load: missing pre-tokenizer type, using:
'default'
load:
load: ************************************
load: GENERATION QUALITY WILL BE
DEGRADED!
load: CONSIDER REGENERATING THE
MODEL
load: ************************************
load:
init_tokenizer: initializing tokenizer for type 2
load: control token: 128255 '<|
reserved_special_token_250|>' is not
```

```
marked as EOG
load: control token: 128253 '<|
reserved_special_token_248|>' is not
marked as EOG
load: control token: 128251 '<|
reserved_special_token_246|>' is not
marked as EOG
load: control token: 128249 '<|
reserved_special_token_244|>' is not
marked as EOG
load: control token: 128248 '<|
reserved_special_token_243|>' is not
marked as EOG
load: control token: 128247 '<|
reserved_special_token_242|>' is not
marked as EOG
load: control token: 128245 '<|
reserved_special_token_240|>' is not
marked as EOG
load: control token: 128244 '<|
reserved_special_token_239|>' is not
marked as EOG
load: control token: 128242 '<|
reserved_special_token_237|>' is not
marked as EOG
```

```
load: control token: 128241 '<|
reserved_special_token_236|>' is not
marked as EOG
load: control token: 128240 '<|
reserved_special_token_235|>' is not
marked as EOG
load: control token: 128237 '<|
reserved_special_token_232|>' is not
marked as EOG
load: control token: 128235 '<|
reserved_special_token_230|>' is not
marked as EOG
load: control token: 128232 '<|
reserved_special_token_227|>' is not
marked as EOG
load: control token: 128231 '<|
reserved_special_token_226|>' is not
marked as EOG
load: control token: 128226 '<|
reserved_special_token_221|>' is not marked
as EOG
load: control token: 128224 '<|
reserved_special_token_219|>' is not
marked as EOG
load: control token: 128223 '<|
reserved_special_token_218|>' is not marked
as EOG
load: control token: 128221 '<|
reserved_special_token_216|>' is not
marked as EOG
load: control token: 128220 '<|
reserved_special_token_215|>' is not marked
as EOG
load: control token: 128218 '<|
reserved_special_token_213|>' is not marked
as EOG
load: control token: 128216 '<|
reserved_special_token_211|>' is not marked
as EOG
load: control token: 128215 '<|
reserved_special_token_210|>' is not marked
as EOG
load: control token: 128214 '<|
reserved_special_token_209|>' is not
marked as EOG
load: control token: 128213 '<|
reserved_special_token_208|>' is not
marked as EOG
load: control token: 128212 '<|
reserved_special_token_207|>' is not
```

```
marked as EOG
load: control token: 128210 '<|
reserved_special_token_205|>' is not
marked as EOG
load: control token: 128208 '<|
reserved_special_token_203|>' is not
marked as EOG
load: control token: 128207 '<|
reserved_special_token_202|>' is not
marked as EOG
load: control token: 128206 '<|
reserved_special_token_201|>' is not marked
as EOG
load: control token: 128205 '<|
reserved_special_token_200|>' is not
marked as EOG
load: control token: 128204 '<|
reserved_special_token_199|>' is not
marked as EOG
load: control token: 128201 '<|
reserved_special_token_196|>' is not
marked as EOG
load: control token: 128199 '<|
reserved_special_token_194|>' is not
marked as EOG
```

```
load: control token: 128194 '<|
reserved_special_token_189|>' is not
marked as EOG
load: control token: 128192 '<|
reserved_special_token_187|>' is not marked
as EOG
load: control token: 128191 '<|
reserved_special_token_186|>' is not
marked as EOG
load: control token: 128188 '<|
reserved_special_token_183|>' is not marked
as EOG
load: control token: 128187 '<|
reserved_special_token_182|>' is not marked
as EOG
load: control token: 128185 '<|
reserved_special_token_180|>' is not
marked as EOG
load: control token: 128184 '<|
reserved_special_token_179|>' is not marked
as EOG
load: control token: 128182 '<|
reserved_special_token_177|>' is not marked
as EOG
load: control token: 128181 '<|
reserved_special_token_176|>' is not marked
as EOG
load: control token: 128180 '<|
reserved_special_token_175|>' is not marked
as EOG
load: control token: 128175 '<|
reserved_special_token_170|>' is not marked
as EOG
load: control token: 128174 '<|
reserved_special_token_169|>' is not
marked as EOG
load: control token: 128173 '<|
reserved_special_token_168|>' is not
marked as EOG
load: control token: 128172 '<|
reserved_special_token_167|>' is not marked
as EOG
load: control token: 128171 '<|
reserved_special_token_166|>' is not
marked as EOG
load: control token: 128170 '<|
reserved_special_token_165|>' is not
marked as EOG
load: control token: 128169 '<|
reserved_special_token_164|>' is not
```

```
marked as EOG
load: control token: 128166 '<|
reserved_special_token_161|>' is not marked
as EOG
load: control token: 128164 '<|
reserved_special_token_159|>' is not
marked as EOG
load: control token: 128163 '<|
reserved_special_token_158|>' is not marked
as EOG
load: control token: 128157 '<|
reserved_special_token_152|>' is not marked
as EOG
load: control token: 128156 '<|
reserved_special_token_151|>' is not marked
as EOG
load: control token: 128154 '<|
reserved_special_token_149|>' is not
marked as EOG
load: control token: 128153 '<|
reserved_special_token_148|>' is not
marked as EOG
load: control token: 128151 '<|
reserved_special_token_146|>' is not
marked as EOG
```

```
load: control token: 128149 '<|
reserved_special_token_144|>' is not
marked as EOG
load: control token: 128148 '<|
reserved_special_token_143|>' is not
marked as EOG
load: control token: 128147 '<|
reserved_special_token_142|>' is not marked
as EOG
load: control token: 128144 '<|
reserved_special_token_139|>' is not
marked as EOG
load: control token: 128141 '<|
reserved_special_token_136|>' is not
marked as EOG
load: control token: 128139 '<|
reserved_special_token_134|>' is not
marked as EOG
load: control token: 128138 '<|
reserved_special_token_133|>' is not marked
as EOG
load: control token: 128137 '<|
reserved_special_token_132|>' is not marked
as EOG
load: control token: 128130 '<|
reserved_special_token_125|>' is not marked
as EOG
load: control token: 128127 '<|
reserved_special_token_122|>' is not marked
as EOG
load: control token: 128125 '<|
reserved_special_token_120|>' is not marked
as EOG
load: control token: 128124 '<|
reserved_special_token_119|>' is not marked
as EOG
load: control token: 128123 '<|
reserved_special_token_118|>' is not marked
as EOG
load: control token: 128122 '<|
reserved_special_token_117|>' is not marked
as EOG
load: control token: 128121 '<|
reserved_special_token_116|>' is not marked
as EOG
load: control token: 128120 '<|
reserved_special_token_115|>' is not marked
as EOG
load: control token: 128119 '<|
reserved_special_token_114|>' is not marked
```

```
as EOG
load: control token: 128118 '<|
reserved_special_token_113|>' is not marked
as EOG
load: control token: 128117 '<|
reserved_special_token_112|>' is not marked
as EOG
load: control token: 128116 '<|
reserved_special_token_111|>' is not marked
as EOG
load: control token: 128113 '<|
reserved_special_token_108|>' is not
marked as EOG
load: control token: 128112 '<|
reserved_special_token_107|>' is not marked
as EOG
load: control token: 128111 '<|
reserved_special_token_106|>' is not
marked as EOG
load: control token: 128110 '<|
reserved_special_token_105|>' is not marked
as EOG
load: control token: 128108 '<|
reserved_special_token_103|>' is not marked
as EOG
```

```
load: control token: 128107 '<|
reserved_special_token_102|>' is not marked
as EOG
load: control token: 128104 '<|
reserved_special_token_99|>' is not marked
as EOG
load: control token: 128103 '<|
reserved_special_token_98|>' is not marked
as EOG
load: control token: 128102 '<|
reserved_special_token_97|>' is not marked
as EOG
load: control token: 128101 '<|
reserved_special_token_96|>' is not marked
as EOG
load: control token: 128100 '<|
reserved_special_token_95|>' is not marked
as EOG
load: control token: 128097 '<|
reserved_special_token_92|>' is not marked
as EOG
load: control token: 128094 '<|
reserved_special_token_89|>' is not marked
as EOG
load: control token: 128093 '<|
reserved_special_token_88|>' is not marked
as EOG
load: control token: 128091 '<|
reserved_special_token_86|>' is not marked
as EOG
load: control token: 128090 '<|
reserved_special_token_85|>' is not marked
as EOG
load: control token: 128087 '<|
reserved_special_token_82|>' is not marked
as EOG
load: control token: 128086 '<|
reserved_special_token_81|>' is not marked
as EOG
load: control token: 128084 '<|
reserved_special_token_79|>' is not marked
as EOG
load: control token: 128082 '<|
reserved_special_token_77|>' is not marked
as EOG
load: control token: 128077 '<|
reserved_special_token_72|>' is not marked
as EOG
load: control token: 128074 '<|
reserved_special_token_69|>' is not marked
```

```
as EOG
load: control token: 128073 '<|
reserved_special_token_68|>' is not marked
as EOG
load: control token: 128070 '<|
reserved_special_token_65|>' is not marked
as EOG
load: control token: 128067 '<|
reserved_special_token_62|>' is not marked
as EOG
load: control token: 128066 '<|
reserved_special_token_61|>' is not marked
as EOG
load: control token: 128064 '<|
reserved_special_token_59|>' is not marked
as EOG
load: control token: 128061 '<|
reserved_special_token_56|>' is not marked
as EOG
load: control token: 128059 '<|
reserved_special_token_54|>' is not marked
as EOG
load: control token: 128058 '<|
reserved_special_token_53|>' is not marked
as EOG
```

```
load: control token: 128057 '<|
reserved_special_token_52|>' is not marked
as EOG
load: control token: 128051 '<|
reserved_special_token_46|>' is not marked
as EOG
load: control token: 128042 '<|
reserved_special_token_37|>' is not marked
as EOG
load: control token: 128041 '<|
reserved_special_token_36|>' is not marked
as EOG
load: control token: 128040 '<|
reserved_special_token_35|>' is not marked
as EOG
load: control token: 128039 '<|
reserved_special_token_34|>' is not marked
as EOG
load: control token: 128035 '<|
reserved_special_token_30|>' is not marked
as EOG
load: control token: 128034 '<|
reserved_special_token_29|>' is not marked
as EOG
load: control token: 128032 '<|
reserved_special_token_27|>' is not marked
as EOG
load: control token: 128031 '<|
reserved_special_token_26|>' is not marked
as EOG
load: control token: 128030 '<|
reserved_special_token_25|>' is not marked
as EOG
load: control token: 128029 '<|
reserved_special_token_24|>' is not marked
as EOG
load: control token: 128027 '<|
reserved_special_token_22|>' is not marked
as EOG
load: control token: 128026 '<|
reserved_special_token_21|>' is not marked
as EOG
load: control token: 128025 '<|
reserved_special_token_20|>' is not marked
as EOG
load: control token: 128023 '<|
reserved_special_token_18|>' is not marked
as EOG
load: control token: 128022 '<|
reserved_special_token_17|>' is not marked
```

```
as EOG
load: control token: 128021 '<|
reserved_special_token_16|>' is not marked
as EOG
load: control token: 128019 '<|
reserved_special_token_14|>' is not marked
as EOG
load: control token: 128017 '<|
reserved_special_token_12|>' is not marked
as EOG
load: control token: 128014 '<|
reserved_special_token_9|>' is not marked
as EOG
load: control token: 128013 '<|
reserved_special_token_8|>' is not marked
as EOG
load: control token: 128012 '<|
reserved_special_token_7|>' is not marked
as EOG
load: control token: 128011 '<|
reserved_special_token_6|>' is not marked
as EOG
load: control token: 128010 '<|
reserved_special_token_5|>' is not marked
as EOG
```

```
load: control token: 128006 '<|
start_header_id|>' is not marked as EOG
load: control token: 128005 '<|
reserved_special_token_3|>' is not marked
as EOG
load: control token: 128003 '<|
reserved_special_token_1|>' is not marked as
EOG
load: control token: 128002 '<|
reserved_special_token_0|>' is not marked
as EOG
load: control token: 128000 '<|begin_of_text|
>' is not marked as EOG
load: control token: 128038 '<|
reserved_special_token_33|>' is not marked
as EOG
load: control token: 128060 '<|
reserved_special_token_55|>' is not marked
as EOG
load: control token: 128043 '<|
reserved_special_token_38|>' is not marked
as EOG
load: control token: 128007 '<|
end_header_id|>' is not marked as EOG
load: control token: 128062 '<|
reserved_special_token_57|>' is not marked
as EOG
load: control token: 128168 '<|
reserved_special_token_163|>' is not
marked as EOG
load: control token: 128159 '<|
reserved_special_token_154|>' is not marked
as EOG
load: control token: 128162 '<|
reserved_special_token_157|>' is not marked
as EOG
load: control token: 128054 '<|
reserved_special_token_49|>' is not marked
as EOG
load: control token: 128047 '<|
reserved_special_token_42|>' is not marked
as EOG
load: control token: 128053 '<|
reserved_special_token_48|>' is not marked
as EOG
load: control token: 128227 '<|
reserved_special_token_222|>' is not
marked as EOG
load: control token: 128095 '<|
reserved_special_token_90|>' is not marked
```

```
as EOG
load: control token: 128150 '<|
reserved_special_token_145|>' is not marked
as EOG
load: control token: 128081 '<|
reserved_special_token_76|>' is not marked
as EOG
load: control token: 128079 '<|
reserved_special_token_74|>' is not marked
as EOG
load: control token: 128099 '<|
reserved_special_token_94|>' is not marked
as EOG
load: control token: 128250 '<|
reserved_special_token_245|>' is not
marked as EOG
load: control token: 128176 '<|
reserved_special_token_171|>' is not marked
as EOG
load: control token: 128068 '<|
reserved_special_token_63|>' is not marked
as EOG
load: control token: 128132 '<|
reserved_special_token_127|>' is not marked
as EOG
```

```
load: control token: 128158 '<|
reserved_special_token_153|>' is not marked
as EOG
load: control token: 128161 '<|
reserved_special_token_156|>' is not
marked as EOG
load: control token: 128131 '<|
reserved_special_token_126|>' is not marked
as EOG
load: control token: 128246 '<|
reserved_special_token_241|>' is not marked
as EOG
load: control token: 128254 '<|
reserved_special_token_249|>' is not
marked as EOG
load: control token: 128033 '<|
reserved_special_token_28|>' is not marked
as EOG
load: control token: 128145 '<|
reserved_special_token_140|>' is not
marked as EOG
load: control token: 128178 '<|
reserved_special_token_173|>' is not marked
as EOG
load: control token: 128219 '<|
reserved_special_token_214|>' is not marked
as EOG
load: control token: 128072 '<|
reserved_special_token_67|>' is not marked
as EOG
load: control token: 128238 '<|
reserved_special_token_233|>' is not
marked as EOG
load: control token: 128048 '<|
reserved_special_token_43|>' is not marked
as EOG
load: control token: 128065 '<|
reserved_special_token_60|>' is not marked
as EOG
load: control token: 128146 '<|
reserved_special_token_141|>' is not marked
as EOG
load: control token: 128198 '<|
reserved_special_token_193|>' is not
marked as EOG
load: control token: 128055 '<|
reserved_special_token_50|>' is not marked
as EOG
load: control token: 128143 '<|
reserved_special_token_138|>' is not marked
```

```
as EOG
load: control token: 128140 '<|
reserved_special_token_135|>' is not marked
as EOG
load: control token: 128020 '<|
reserved_special_token_15|>' is not marked
as EOG
load: control token: 128036 '<|
reserved_special_token_31|>' is not marked
as EOG
load: control token: 128129 '<|
reserved_special_token_124|>' is not marked
as EOG
load: control token: 128098 '<|
reserved_special_token_93|>' is not marked
as EOG
load: control token: 128209 '<|
reserved_special_token_204|>' is not
marked as EOG
load: control token: 128186 '<|
reserved_special_token_181|>' is not marked
as EOG
load: control token: 128222 '<|
reserved_special_token_217|>' is not marked
as EOG
```

```
load: control token: 128126 '<|
reserved_special_token_121|>' is not marked
as EOG
load: control token: 128004 '<|
reserved_special_token_2|>' is not marked
as EOG
load: control token: 128075 '<|
reserved_special_token_70|>' is not marked
as EOG
load: control token: 128160 '<|
reserved_special_token_155|>' is not marked
as EOG
load: control token: 128069 '<|
reserved_special_token_64|>' is not marked
as EOG
load: control token: 128109 '<|
reserved_special_token_104|>' is not
marked as EOG
load: control token: 128183 '<|
reserved_special_token_178|>' is not marked
as EOG
load: control token: 128092 '<|
reserved_special_token_87|>' is not marked
as EOG
load: control token: 128106 '<|
reserved_special_token_101|>' is not marked
as EOG
load: control token: 128096 '<|
reserved_special_token_91|>' is not marked
as EOG
load: control token: 128135 '<|
reserved_special_token_130|>' is not marked
as EOG
load: control token: 128190 '<|
reserved_special_token_185|>' is not marked
as EOG
load: control token: 128196 '<|
reserved_special_token_191|>' is not marked
as EOG
load: control token: 128045 '<|
reserved_special_token_40|>' is not marked
as EOG
load: control token: 128085 '<|
reserved_special_token_80|>' is not marked
as EOG
load: control token: 128189 '<|
reserved_special_token_184|>' is not
marked as EOG
load: control token: 128133 '<|
reserved_special_token_128|>' is not marked
```

```
as EOG
load: control token: 128089 '<|
reserved_special_token_84|>' is not marked
as EOG
load: control token: 128155 '<|
reserved_special_token_150|>' is not marked
as EOG
load: control token: 128046 '<|
reserved_special_token_41|>' is not marked
as EOG
load: control token: 128028 '<|
reserved_special_token_23|>' is not marked
as EOG
load: control token: 128252 '<|
reserved_special_token_247|>' is not
marked as EOG
load: control token: 128179 '<|
reserved_special_token_174|>' is not marked
as EOG
load: control token: 128063 '<|
reserved_special_token_58|>' is not marked
as EOG
load: control token: 128177 '<|
reserved_special_token_172|>' is not marked
as EOG
```

```
load: control token: 128230 '<|
reserved_special_token_225|>' is not
marked as EOG
load: control token: 128076 '<|
reserved_special_token_71|>' is not marked
as EOG
load: control token: 128078 '<|
reserved_special_token_73|>' is not marked
as EOG
load: control token: 128228 '<|
reserved_special_token_223|>' is not
marked as EOG
load: control token: 128193 '<|
reserved_special_token_188|>' is not
marked as EOG
load: control token: 128044 '<|
reserved_special_token_39|>' is not marked
as EOG
load: control token: 128080 '<|
reserved_special_token_75|>' is not marked
as EOG
load: control token: 128136 '<|
reserved_special_token_131|>' is not marked
as EOG
load: control token: 128128 '<|
reserved_special_token_123|>' is not marked
as EOG
load: control token: 128115 '<|
reserved_special_token_110|>' is not marked
as EOG
load: control token: 128050 '<|
reserved_special_token_45|>' is not marked
as EOG
load: control token: 128217 '<|
reserved_special_token_212|>' is not marked
as EOG
load: control token: 128105 '<|
reserved_special_token_100|>' is not
marked as EOG
load: control token: 128088 '<|
reserved_special_token_83|>' is not marked
as EOG
load: control token: 128200 '<|
reserved_special_token_195|>' is not
marked as EOG
load: control token: 128056 '<|
reserved_special_token_51|>' is not marked
as EOG
load: control token: 128016 '<|
reserved_special_token_11|>' is not marked
```

```
as EOG
load: control token: 128167 '<|
reserved_special_token_162|>' is not marked
as EOG
load: control token: 128202 '<|
reserved_special_token_197|>' is not marked
as EOG
load: control token: 128037 '<|
reserved_special_token_32|>' is not marked
as EOG
load: control token: 128197 '<|
reserved_special_token_192|>' is not marked
as EOG
load: control token: 128233 '<|
reserved_special_token_228|>' is not
marked as EOG
load: control token: 128142 '<|
reserved_special_token_137|>' is not marked
as EOG
load: control token: 128165 '<|
reserved_special_token_160|>' is not
marked as EOG
load: control token: 128211 '<|
reserved_special_token_206|>' is not
marked as EOG
```

```
load: control token: 128134 '<|
reserved_special_token_129|>' is not marked
as EOG
load: control token: 128229 '<|
reserved_special_token_224|>' is not
marked as EOG
load: control token: 128236 '<|
reserved_special_token_231|>' is not marked
as EOG
load: control token: 128052 '<|
reserved_special_token_47|>' is not marked
as EOG
load: control token: 128225 '<|
reserved_special_token_220|>' is not
marked as EOG
load: control token: 128203 '<|
reserved_special_token_198|>' is not
marked as EOG
load: control token: 128015 '<|
reserved_special_token_10|>' is not marked
as EOG
load: control token: 128008 '<|
reserved_special_token_4|>' is not marked
as EOG
load: control token: 128195 '<|
reserved_special_token_190|>' is not
marked as EOG
load: control token: 128018 '<|
reserved_special_token_13|>' is not marked
as EOG
load: control token: 128083 '<|
reserved_special_token_78|>' is not marked
as EOG
load: control token: 128071 '<|
reserved_special_token_66|>' is not marked
as EOG
load: control token: 128024 '<|
reserved_special_token_19|>' is not marked
as EOG
load: control token: 128239 '<|
reserved_special_token_234|>' is not
marked as EOG
load: control token: 128152 '<|
reserved_special_token_147|>' is not marked
as EOG
load: control token: 128049 '<|
reserved_special_token_44|>' is not marked
as EOG
load: control token: 128243 '<|
reserved_special_token_238|>' is not
```

```
marked as EOG
load: control token: 128114 '<|
reserved_special_token_109|>' is not
marked as EOG
load: control token: 128234 '<|
reserved_special_token_229|>' is not
marked as EOG
load: special tokens cache size = 256
load: token to piece cache size = 0.8000 MB
print_info: arch             = llama
print_info: vocab_only       = 0
print_info: n_ctx_train      = 8192
print_info: n_embd           = 4096
print_info: n_layer          = 32
print_info: n_head           = 32
print_info: n_head_kv        = 8
print_info: n_rot            = 128
print_info: n_swa            = 0
print_info: is_swa_any       = 0
print_info: n_embd_head_k    = 128
print_info: n_embd_head_v    = 128
print_info: n_gqa            = 4
print_info: n_embd_k_gqa     = 1024
print_info: n_embd_v_gqa     = 1024
print_info: f_norm_eps       = 0.0e+00
```

```
print_info: f_norm_rms_eps   = 1.0e-05
print_info: f_clamp_kqv      = 0.0e+00
print_info: f_max_alibi_bias = 0.0e+00
print_info: f_logit_scale    = 0.0e+00
print_info: f_attn_scale     = 0.0e+00
print_info: n_ff             = 14336
print_info: n_expert         = 0
print_info: n_expert_used    = 0
print_info: causal attn      = 1
print_info: pooling type     = 0
print_info: rope type        = 0
print_info: rope scaling     = linear
print_info: freq_base_train  = 500000.0
print_info: freq_scale_train = 1
print_info: n_ctx_orig_yarn  = 8192
print_info: rope_finetuned   = unknown
print_info: model type       = 8B
print_info: model params     = 8.03 B
print_info: general.name     = Meta-
Llama-3-8B-Instruct
print_info: vocab type       = BPE
print_info: n_vocab          = 128256
print_info: n_merges         = 280147
print_info: BOS token        = 128000 '<|
begin_of_text|>'
```

```
print_info: EOS token        = 128001 '<|
end_of_text|>'
print_info: EOT token        = 128009 '<|eot_id|
>'
print_info: LF token         = 198 'Ċ'
print_info: EOG token        = 128001 '<|
end_of_text|>'
print_info: EOG token        = 128009 '<|
eot_id|>'
print_info: max token length = 256
load_tensors: loading model tensors, this can
take a while... (mmap = true)
load_tensors: layer   0 assigned to device
Metal, is_swa = 0
load_tensors: layer   1 assigned to device
Metal, is_swa = 0
load_tensors: layer   2 assigned to device
Metal, is_swa = 0
load_tensors: layer   3 assigned to device
Metal, is_swa = 0
load_tensors: layer   4 assigned to device
Metal, is_swa = 0
load_tensors: layer   5 assigned to device
Metal, is_swa = 0
load_tensors: layer   6 assigned to device
```

```
Metal, is_swa = 0
load_tensors: layer   7 assigned to device
Metal, is_swa = 0
load_tensors: layer   8 assigned to device
Metal, is_swa = 0
load_tensors: layer   9 assigned to device
Metal, is_swa = 0
load_tensors: layer  10 assigned to device
Metal, is_swa = 0
load_tensors: layer  11 assigned to device
Metal, is_swa = 0
load_tensors: layer  12 assigned to device
Metal, is_swa = 0
load_tensors: layer  13 assigned to device
Metal, is_swa = 0
load_tensors: layer  14 assigned to device
Metal, is_swa = 0
load_tensors: layer  15 assigned to device
Metal, is_swa = 0
load_tensors: layer  16 assigned to device
Metal, is_swa = 0
load_tensors: layer  17 assigned to device
Metal, is_swa = 0
load_tensors: layer  18 assigned to device
Metal, is_swa = 0
```

```
load_tensors: layer  19 assigned to device
Metal, is_swa = 0
load_tensors: layer  20 assigned to device
Metal, is_swa = 0
load_tensors: layer  21 assigned to device
Metal, is_swa = 0
load_tensors: layer  22 assigned to device
Metal, is_swa = 0
load_tensors: layer  23 assigned to device
Metal, is_swa = 0
load_tensors: layer  24 assigned to device
Metal, is_swa = 0
load_tensors: layer  25 assigned to device
Metal, is_swa = 0
load_tensors: layer  26 assigned to device
Metal, is_swa = 0
load_tensors: layer  27 assigned to device
Metal, is_swa = 0
load_tensors: layer  28 assigned to device
Metal, is_swa = 0
load_tensors: layer  29 assigned to device
Metal, is_swa = 0
load_tensors: layer  30 assigned to device
Metal, is_swa = 0
load_tensors: layer  31 assigned to device
```

```
Metal, is_swa = 0
load_tensors: layer  32 assigned to device
Metal, is_swa = 0
load_tensors: tensor 'token_embd.weight'
(q4_K) (and 0 others) cannot be used with
preferred buffer type CPU_REPACK, using
CPU instead
ggml_backend_metal_log_allocated_size:
allocated buffer, size =  4403.50 MiB,
( 4403.56 / 10922.67)
load_tensors: offloading 32 repeating layers
to GPU
load_tensors: offloading output layer to GPU
load_tensors: offloaded 33/33 layers to GPU
load_tensors: Metal_Mapped model buffer
size =  4403.50 MiB
load_tensors:   CPU_Mapped model buffer
size =   281.81 MiB
..............................................................................
....
llama_context: constructing llama_context
llama_context: n_seq_max     = 1
llama_context: n_ctx         = 4096
llama_context: n_ctx_per_seq = 4096
llama_context: n_batch       = 512
```

```
llama_context: n_ubatch      = 512
llama_context: causal_attn   = 1
llama_context: flash_attn    = 0
llama_context: freq_base     = 500000.0
llama_context: freq_scale    = 1
llama_context: n_ctx_per_seq (4096) <
n_ctx_train (8192) -- the full capacity of the
model will not be utilized
ggml_metal_init: allocating
ggml_metal_init: found device: Apple M1 Pro
ggml_metal_init: picking default device:
Apple M1 Pro
ggml_metal_load_library: using embedded
metal library
ggml_metal_init: GPU name:   Apple M1 Pro
ggml_metal_init: GPU family:
MTLGPUFamilyApple7  (1007)
ggml_metal_init: GPU family:
MTLGPUFamilyCommon3 (3003)
ggml_metal_init: GPU family:
MTLGPUFamilyMetal3  (5001)
ggml_metal_init: simdgroup reduction   = true
ggml_metal_init: simdgroup matrix mul. =
true
ggml_metal_init: has residency sets    = true
```

```
ggml_metal_init: has bfloat            = true
ggml_metal_init: use bfloat            = false
ggml_metal_init: hasUnifiedMemory      = true
ggml_metal_init:
recommendedMaxWorkingSetSize  =
11453.25 MB
ggml_metal_init: loaded kernel_add
0x10783cc80 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_add_row
0x10783d0e0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_sub
0x10783d800 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_sub_row
0x10783df20 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_mul
0x10783e640 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_mul_row
0x126604400 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_div
```

```
0x10783f160 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_div_row
0x10669acd0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_repeat_f32
0x1079152f0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_repeat_f16
0x1267be7c0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_repeat_i32
0x10783f710 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_repeat_i16
0x1267bedd0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_scale
0x1267bf3a0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_scale_4
0x10783fe10 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_clamp
0x107840940 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_tanh
0x1066a0a40 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_relu
0x107840e20 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_sigmoid
0x1266c1350 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_gelu
0x1078414d0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_gelu_4
0x107841fc0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_gelu_erf
0x1267c02b0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_gelu_erf_4
0x1078427d0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_gelu_quick
0x107843030 | th_max = 1024 | th_width =
32
```

```
ggml_metal_init: loaded kernel_gelu_quick_4
0x1267c0aa0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_silu
0x1267c1190 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_silu_4
0x107843840 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_elu
0x1078440a0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_abs
0x1267c1880 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_sgn
0x1078448b0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_step
0x107845110 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_hardswish
0x107845920 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_hardsigmoid
```

```
0x107846130 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_exp
0x1078466c0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_soft_max_f16
0x1267c1ef0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_soft_max_f16_4
0x1267c2330 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_soft_max_f32
0x1267c2a50 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_soft_max_f32_4
0x107846c50 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_diag_mask_inf
0x107847090 | th_max = 1024 | th_width =
32
```

```
ggml_metal_init: loaded
kernel_diag_mask_inf_8
0x1078476b0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_get_rows_f32
0x107847ce0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_get_rows_f16
0x1267c34c0 | th_max = 1024 | th_width =
32
ggml_metal_init: skipping
kernel_get_rows_bf16                 (not
supported)
ggml_metal_init: loaded
kernel_get_rows_q4_0
0x1267c3720 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_get_rows_q4_1
0x107848650 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_get_rows_q5_0
```

```
0x1267c4090 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_get_rows_q5_1
0x107848c00 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_get_rows_q8_0
0x1267c4640 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_get_rows_q2_K
0x1078491b0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_get_rows_q3_K
0x107849e00 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_get_rows_q4_K
0x1267c4a10 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_get_rows_q5_K
0x1267c4e50 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_get_rows_q6_K
0x10784a240 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_get_rows_iq2_xxs
0x10784a4a0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_get_rows_iq2_xs
0x1267c5290 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_get_rows_iq3_xxs
0x10784ae30 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_get_rows_iq3_s
0x10784b090 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_get_rows_iq2_s
0x104e09c60 | th_max = 1024 | th_width =
32
```

```
ggml_metal_init: loaded
kernel_get_rows_iq1_s
0x10784ba90 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_get_rows_iq1_m
0x1267c5850 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_get_rows_iq4_nl
0x10784c0d0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_get_rows_iq4_xs
0x1267c1ae0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_get_rows_i32
0x1267c5d80 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_set_rows_f32
0x1267c6780 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_set_rows_f16
```

```
0x10784c690 | th_max = 1024 | th_width =
32
ggml_metal_init: skipping
kernel_set_rows_bf16                    (not
supported)
ggml_metal_init: loaded
kernel_set_rows_q8_0
0x1267c6bc0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_set_rows_q4_0
0x10784cc40 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_set_rows_q4_1
0x1267c7170 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_set_rows_q5_0
0x10784d1f0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_set_rows_q5_1
0x10784d910 | th_max = 1024 | th_width =
32
```

```
ggml_metal_init: loaded
kernel_set_rows_iq4_nl
0x10784db70 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_rms_norm
0x1267c7720 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_l2_norm
0x10784e3d0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_group_norm
0x10784e630 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_norm
0x10784ec50 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_ssm_conv_f32
0x10784f730 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_ssm_scan_f32
0x10784fd90 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
```

```
kernel_ssm_scan_f32_group
0x10784fff0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_rwkv_wkv6_f32
0x107850400 | th_max =  384 | th_width =
32
ggml_metal_init: loaded
kernel_rwkv_wkv7_f32
0x107850a20 | th_max =  448 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_f32_f32
0x1267c7cd0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_f32_f32_c4
0x107851390 | th_max = 1024 | th_width =
32
ggml_metal_init: skipping
kernel_mul_mv_bf16_f32               (not
supported)
ggml_metal_init: skipping
kernel_mul_mv_bf16_f32_c4             (not
supported)
```

```
ggml_metal_init: skipping
kernel_mul_mv_bf16_f32_1row            (not
supported)
ggml_metal_init: skipping
kernel_mul_mv_bf16_f32_l4              (not
supported)
ggml_metal_init: skipping
kernel_mul_mv_bf16_bf16                (not
supported)
ggml_metal_init: loaded
kernel_mul_mv_f16_f32
0x107851ab0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_f16_f32_c4
0x107851d10 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_f16_f32_1row
0x1078526f0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_f16_f32_l4
0x107852ca0 | th_max = 1024 | th_width =
32
```

```
ggml_metal_init: loaded
kernel_mul_mv_f16_f16
0x1267c8290 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_q4_0_f32
0x1267c89c0 | th_max =  640 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_q4_1_f32
0x1078530e0 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_q5_0_f32
0x10794b6a0 | th_max =  640 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_q5_1_f32
0x1078536a0 | th_max =  576 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_q8_0_f32
0x107853dd0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
```

```
kernel_mul_mv_ext_f16_f32_r1_2
0x1267c8c20 | th_max =  896 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_f16_f32_r1_3
0x107854030 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_f16_f32_r1_4
0x107854440 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_f16_f32_r1_5
0x107854e30 | th_max =  768 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_q4_0_f32_r1_2
0x1078555b0 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_q4_0_f32_r1_3
0x107855930 | th_max =  704 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_q4_0_f32_r1_4
```

```
0x107855e30 | th_max =  640 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_q4_0_f32_r1_5
0x1267c9360 | th_max =  640 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_q4_1_f32_r1_2
0x107856090 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_q4_1_f32_r1_3
0x1267c9ad0 | th_max =  704 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_q4_1_f32_r1_4
0x1078562f0 | th_max =  640 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_q4_1_f32_r1_5
0x1267ca230 | th_max =  640 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_q5_0_f32_r1_2
0x1267ca490 | th_max =  704 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_q5_0_f32_r1_3
0x1078568a0 | th_max =  704 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_q5_0_f32_r1_4
0x107856cb0 | th_max =  576 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_q5_0_f32_r1_5
0x1078576a0 | th_max =  640 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_q5_1_f32_r1_2
0x107857e20 | th_max =  640 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_q5_1_f32_r1_3
0x1267caf00 | th_max =  704 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_q5_1_f32_r1_4
0x10794b900 | th_max =  576 | th_width =
32
```

```
ggml_metal_init: loaded
kernel_mul_mv_ext_q5_1_f32_r1_5
0x107858080 | th_max =  640 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_q8_0_f32_r1_2
0x107858760 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_q8_0_f32_r1_3
0x107858cc0 | th_max =  704 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_q8_0_f32_r1_4
0x107859220 | th_max =  640 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_q8_0_f32_r1_5
0x1267cb160 | th_max =  640 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_q4_K_f32_r1_2
0x1267cb9f0 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
```

```
kernel_mul_mv_ext_q4_K_f32_r1_3
0x1267cbf50 | th_max =  704 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_q4_K_f32_r1_4
0x1078595a0 | th_max =  640 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_q4_K_f32_r1_5
0x107859800 | th_max =  640 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_q5_K_f32_r1_2
0x107859e30 | th_max =  704 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_q5_K_f32_r1_3
0x10785a4b0 | th_max =  704 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_q5_K_f32_r1_4
0x1266c15b0 | th_max =  640 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_q5_K_f32_r1_5
```

```
0x10785adf0 | th_max =  640 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_q6_K_f32_r1_2
0x1267cc450 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_q6_K_f32_r1_3
0x1267ccbb0 | th_max =  704 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_q6_K_f32_r1_4
0x10785b170 | th_max =  704 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_q6_K_f32_r1_5
0x1267cce10 | th_max =  640 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_iq4_nl_f32_r1_2
0x10785b6d0 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_iq4_nl_f32_r1_3
0x1267cd580 | th_max =  704 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_iq4_nl_f32_r1_4
0x10785be40 | th_max =  640 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_ext_iq4_nl_f32_r1_5
0x10785c5a0 | th_max =  640 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_q2_K_f32
0x10785c920 | th_max =  640 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_q3_K_f32
0x1267cda50 | th_max =  576 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_q4_K_f32
0x10785cca0 | th_max =  576 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_q5_K_f32
0x10785d3d0 | th_max =  576 | th_width =
32
```

```
ggml_metal_init: loaded
kernel_mul_mv_q6_K_f32
0x10785d980 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_iq2_xxs_f32
0x1267cdcb0 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_iq2_xs_f32
0x10794c890 | th_max =  704 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_iq3_xxs_f32
0x1267ce520 | th_max =  768 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_iq3_s_f32
0x1266c1b10 | th_max =  640 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_iq2_s_f32
0x10785ddc0 | th_max =  704 | th_width =
32
ggml_metal_init: loaded
```

```
kernel_mul_mv_iq1_s_f32
0x10785e020 | th_max =  448 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_iq1_m_f32
0x1267ceae0 | th_max =  576 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_iq4_nl_f32
0x10785e950 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_iq4_xs_f32
0x10785f240 | th_max =  896 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_id_f32_f32
0x10785f7f0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_id_f16_f32
0x1267cf0a0 | th_max = 1024 | th_width =
32
ggml_metal_init: skipping
kernel_mul_mv_id_bf16_f32              (not
supported)
ggml_metal_init: loaded
kernel_mul_mv_id_q4_0_f32
0x10785fc30 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_id_q4_1_f32
0x1267cf660 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_id_q5_0_f32
0x1078601f0 | th_max =  576 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_id_q5_1_f32
0x107860920 | th_max =  576 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_id_q8_0_f32
0x10794cca0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_id_q2_K_f32
0x107e043d0 | th_max =  576 | th_width =
32
```

```
ggml_metal_init: loaded
kernel_mul_mv_id_q3_K_f32
0x106695df0 | th_max =  576 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_id_q4_K_f32
0x10794d0b0 | th_max =  576 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_id_q5_K_f32
0x107860f10 | th_max =  576 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_id_q6_K_f32
0x1066a0cf0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_id_iq2_xxs_f32
0x1078614d0 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_id_iq2_xs_f32
0x107861c00 | th_max =  640 | th_width =
32
ggml_metal_init: loaded
```

```
kernel_mul_mv_id_iq3_xxs_f32
0x1078621b0 | th_max =  704 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_id_iq3_s_f32
0x1266c2070 | th_max =  640 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_id_iq2_s_f32
0x1066a0f50 | th_max =  640 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_id_iq1_s_f32
0x1078625f0 | th_max =  448 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_id_iq1_m_f32
0x107862850 | th_max =  576 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_id_iq4_nl_f32
0x107863260 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mv_id_iq4_xs_f32
```

```
0x107863810 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_f32_f32
0x1267cfbf0 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_f16_f32
0x107863c50 | th_max =  832 | th_width =
32
ggml_metal_init: skipping
kernel_mul_mm_bf16_f32                  (not
supported)
ggml_metal_init: loaded
kernel_mul_mm_q4_0_f32
0x107864380 | th_max =  768 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_q4_1_f32
0x107864930 | th_max =  768 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_q5_0_f32
0x1267d0360 | th_max =  768 | th_width =
32
```

```
ggml_metal_init: loaded
kernel_mul_mm_q5_1_f32
0x107864d70 | th_max =  768 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_q8_0_f32
0x1267d0920 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_q2_K_f32
0x107865330 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_q3_K_f32
0x1267d0ee0 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_q4_K_f32
0x1078658f0 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_q5_K_f32
0x107866020 | th_max =  768 | th_width =
32
ggml_metal_init: loaded
```

```
kernel_mul_mm_q6_K_f32
0x107866460 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_iq2_xxs_f32
0x1267d14a0 | th_max =  768 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_iq2_xs_f32
0x1267d1700 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_iq3_xxs_f32
0x107866ab0 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_iq3_s_f32
0x107866d10 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_iq2_s_f32
0x1267d1fa0 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_iq1_s_f32
```

```
0x1078673d0 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_iq1_m_f32
0x1267d25f0 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_iq4_nl_f32
0x107867be0 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_iq4_xs_f32
0x1267d2bb0 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_id_map0_f16
0x1266c2750 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_id_map1_f32
0x1078681a0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_id_f32_f16
0x1078688d0 | th_max =  896 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_id_f16_f16
0x1267d3170 | th_max =  896 | th_width =
32
ggml_metal_init: skipping
kernel_mul_mm_id_bf16_f16             (not
supported)
ggml_metal_init: loaded
kernel_mul_mm_id_q4_0_f16
0x1267d38a0 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_id_q4_1_f16
0x107868d10 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_id_q5_0_f16
0x1267d3ce0 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_id_q5_1_f16
0x1078692d0 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
```

```
kernel_mul_mm_id_q8_0_f16
0x1267d42a0 | th_max =  896 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_id_q2_K_f16
0x107869890 | th_max =  896 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_id_q3_K_f16
0x1267d4860 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_id_q4_K_f16
0x1267d4f90 | th_max =  896 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_id_q5_K_f16
0x107869e50 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_id_q6_K_f16
0x1267d53d0 | th_max =  896 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_id_iq2_xxs_f16
```

```
0x1267d5630 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_id_iq2_xs_f16
0x10786a410 | th_max =  896 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_id_iq3_xxs_f16
0x1267d5cf0 | th_max =  896 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_id_iq3_s_f16
0x1267d6110 | th_max =  896 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_id_iq2_s_f16
0x10786aa80 | th_max =  896 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_id_iq1_s_f16
0x10786ace0 | th_max =  896 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_id_iq1_m_f16
0x1267d67d0 | th_max =  896 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_id_iq4_nl_f16
0x1266c2cb0 | th_max =  896 | th_width =
32
ggml_metal_init: loaded
kernel_mul_mm_id_iq4_xs_f16
0x10786b3a0 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_rope_norm_f32
0x1267d7000 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_rope_norm_f16
0x1267d74f0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_rope_multi_f32
0x10786bbc0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_rope_multi_f16
0x1267d7750 | th_max = 1024 | th_width =
32
```

```
ggml_metal_init: loaded
kernel_rope_vision_f32
0x10794d970 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_rope_vision_f16
0x1267d7d10 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_rope_neox_f32
0x10786c220 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_rope_neox_f16
0x10786c480 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_im2col_f16
0x10786d060 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_im2col_f32
0x1267d8150 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_im2col_ext_f16
0x1267d8870 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_im2col_ext_f32
0x1267d8f40 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_conv_transpose_1d_f32_f32
0x10786d4a0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_conv_transpose_1d_f16_f32
0x10786d700 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_upscale_f32
0x10786dea0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_pad_f32
0x10786e600 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_pad_reflect_1d_f32
0x10786e860 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_timestep_embedding_f32
```

```
0x10786ed70 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_arange_f32
0x10786f3a0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_argsort_f32_i32_asc
0x10794e080 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_argsort_f32_i32_desc
0x1267d91a0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_leaky_relu_f32
0x1267d9fc0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_f16_h64
0x10786fdd0 | th_max =  704 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_f16_h80
0x107870030 | th_max =  704 | th_width =
32
```

```
ggml_metal_init: loaded
kernel_flash_attn_ext_f16_h96
0x107870760 | th_max =  640 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_f16_h112
0x1267da220 | th_max =  576 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_f16_h128
0x1267da630 | th_max =  576 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_f16_h192
0x1078712a0 | th_max =  576 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_f16_hk192_hv128
0x1267db170 | th_max =  576 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_f16_h256
0x1267db3d0 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
```

```
kernel_flash_attn_ext_f16_hk576_hv512
0x107871680 | th_max =  832 | th_width =
32
ggml_metal_init: skipping
kernel_flash_attn_ext_bf16_h64           (not
supported)
ggml_metal_init: skipping
kernel_flash_attn_ext_bf16_h80           (not
supported)
ggml_metal_init: skipping
kernel_flash_attn_ext_bf16_h96           (not
supported)
ggml_metal_init: skipping
kernel_flash_attn_ext_bf16_h112          (not
supported)
ggml_metal_init: skipping
kernel_flash_attn_ext_bf16_h128          (not
supported)
ggml_metal_init: skipping
kernel_flash_attn_ext_bf16_h192          (not
supported)
ggml_metal_init: skipping
kernel_flash_attn_ext_bf16_hk192_hv128
(not supported)
ggml_metal_init: skipping
```

```
kernel_flash_attn_ext_bf16_h256 (not
supported)
ggml_metal_init: skipping
kernel_flash_attn_ext_bf16_hk576_hv512
(not supported)
ggml_metal_init: loaded
kernel_flash_attn_ext_q4_0_h64
0x1267dbf10 | th_max =  640 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q4_0_h80
0x1267dc170 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q4_0_h96
0x107871a90 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q4_0_h112
0x107871ea0 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q4_0_h128
0x1267dcbe0 | th_max =  768 | th_width =
32
```

```
ggml_metal_init: loaded
kernel_flash_attn_ext_q4_0_h192
0x1267dce40 | th_max =  768 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q4_0_hk192_hv128
0x1078383d0 | th_max =  768 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q4_0_h256
0x107872730 | th_max =  768 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q4_0_hk576_hv512
0x1267dd730 | th_max =  768 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q4_1_h64
0x1078731b0 | th_max =  704 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q4_1_h80
0x1267ddab0 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
```

```
kernel_flash_attn_ext_q4_1_h96
0x107873410 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q4_1_h112
0x107873ca0 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q4_1_h128
0x107874200 | th_max =  768 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q4_1_h192
0x107874b00 | th_max =  768 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q4_1_hk192_hv128
0x1267ddd10 | th_max =  768 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q4_1_h256
0x107874d60 | th_max =  768 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q4_1_hk576_hv512
```

```
0x1267de2b0 | th_max =  768 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q5_0_h64
0x107875150 | th_max =  576 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q5_0_h80
0x107875680 | th_max =  704 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q5_0_h96
0x107875f10 | th_max =  704 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q5_0_h112
0x107876470 | th_max =  704 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q5_0_h128
0x1267deb60 | th_max =  704 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q5_0_h192
0x1267dedc0 | th_max =  704 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q5_0_hk192_hv128
0x1078766d0 | th_max =  704 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q5_0_h256
0x1267df460 | th_max =  704 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q5_0_hk576_hv512
0x107876ff0 | th_max =  704 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q5_1_h64
0x107877750 | th_max =  576 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q5_1_h80
0x107877ad0 | th_max =  704 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q5_1_h96
0x107877fd0 | th_max =  704 | th_width =
32
```

```
ggml_metal_init: loaded
kernel_flash_attn_ext_q5_1_h112
0x1267dfa30 | th_max =  704 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q5_1_h128
0x107878230 | th_max =  640 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q5_1_h192
0x107878490 | th_max =  640 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q5_1_hk192_hv128
0x107878ac0 | th_max =  640 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q5_1_h256
0x107879830 | th_max =  640 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q5_1_hk576_hv512
0x107879a90 | th_max =  640 | th_width =
32
ggml_metal_init: loaded
```

```
kernel_flash_attn_ext_q8_0_h64
0x1267dfe40 | th_max =  704 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q8_0_h80
0x10787a200 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q8_0_h96
0x10787a960 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q8_0_h112
0x10787ace0 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q8_0_h128
0x1267e0720 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q8_0_h192
0x10787b060 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q8_0_hk192_hv128
```

```
0x1267e0e80 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q8_0_h256
0x1266c3450 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_q8_0_hk576_hv512
0x10787b3e0 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_f16_h64
0x10787b640 | th_max =  896 | th_width =
32
ggml_metal_init: skipping
kernel_flash_attn_ext_vec_bf16_h64      (not
supported)
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_q4_0_h64
0x10787bbb0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_q4_1_h64
0x10787c130 | th_max =  896 | th_width =
32
```

```
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_q5_0_h64
0x10787cbb0 | th_max =  704 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_q5_1_h64
0x1266c39b0 | th_max =  704 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_q8_0_h64
0x10794e700 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_f16_h96
0x10787d220 | th_max = 1024 | th_width =
32
ggml_metal_init: skipping
kernel_flash_attn_ext_vec_bf16_h96      (not
supported)
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_q4_0_h96
0x1267e1200 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_q4_1_h96
```

```
0x10787d890 | th_max =  896 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_q5_0_h96
0x1267e1870 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_q5_1_h96
0x1267e1ad0 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_q8_0_h96
0x104e0aab0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_f16_h128
0x1267e2610 | th_max =  896 | th_width =
32
ggml_metal_init: skipping
kernel_flash_attn_ext_vec_bf16_h128
(not supported)
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_q4_0_h128
0x1267e2870 | th_max =  896 | th_width =
32
```

```
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_q4_1_h128
0x10794ec60 | th_max =  832 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_q5_0_h128
0x10787df00 | th_max =  768 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_q5_1_h128
0x10787e160 | th_max =  768 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_q8_0_h128
0x10794f2f0 | th_max =  896 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_f16_h192
0x1267e33b0 | th_max =  640 | th_width =
32
ggml_metal_init: skipping
kernel_flash_attn_ext_vec_bf16_h192
(not supported)
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_q4_0_h192
```

```
0x10794f550 | th_max =  768 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_q4_1_h192
0x10787eca0 | th_max =  768 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_q5_0_h192
0x10787f680 | th_max =  640 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_q5_1_h192
0x10787fce0 | th_max =  576 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_q8_0_h192
0x107880340 | th_max =  768 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_f16_hk192_hv12
8     0x1267e3730 | th_max =  768 | th_width
=   32
ggml_metal_init: skipping
kernel_flash_attn_ext_vec_bf16_hk192_hv1
28 (not supported)
```

```
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_q4_0_hk192_hv1
28     0x1267e3990 | th_max =  832 |
th_width =   32
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_q4_1_hk192_hv1
28     0x1078805a0 | th_max =  832 |
th_width =   32
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_q5_0_hk192_hv1
28     0x1078809b0 | th_max =  768 |
th_width =   32
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_q5_1_hk192_hv1
28     0x1267e4160 | th_max =  704 |
th_width =   32
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_q8_0_hk192_hv1
28     0x1266c3ee0 | th_max =  832 |
th_width =   32
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_f16_h256
0x1266c4750 | th_max =  576 | th_width =
32
ggml_metal_init: skipping
```

```
kernel_flash_attn_ext_vec_bf16_h256
(not supported)
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_q4_0_h256
0x1267e49c0 | th_max =  640 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_q4_1_h256
0x10787f020 | th_max =  576 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_q5_0_h256
0x1078812b0 | th_max =  512 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_q5_1_h256
0x1078818e0 | th_max =  512 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_q8_0_h256
0x107882050 | th_max =  576 | th_width =
32
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_f16_hk576_hv51
2     0x1266c4ad0 | th_max =  576 | th_width
```

```
=   32
ggml_metal_init: skipping
kernel_flash_attn_ext_vec_bf16_hk576_hv5
12 (not supported)
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_q4_0_hk576_hv
512      0x107882760 | th_max =  576 |
th_width =   32
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_q4_1_hk576_hv5
12      0x1078829c0 | th_max =  576 |
th_width =   32
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_q5_0_hk576_hv
512      0x107883130 | th_max =  512 |
th_width =   32
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_q5_1_hk576_hv5
12      0x107883660 | th_max =  512 |
th_width =   32
ggml_metal_init: loaded
kernel_flash_attn_ext_vec_q8_0_hk576_hv
512      0x1267e4c20 | th_max =  576 |
th_width =   32
ggml_metal_init: loaded kernel_set_f32
```

```
0x1267e54f0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_set_i32
0x1267e5750 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_cpy_f32_f32
0x107883c00 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_cpy_f32_f16
0x107884320 | th_max = 1024 | th_width =
32
ggml_metal_init: skipping
kernel_cpy_f32_bf16                    (not
supported)
ggml_metal_init: loaded kernel_cpy_f16_f32
0x1267e5d00 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_cpy_f16_f16
0x1267e6420 | th_max = 1024 | th_width =
32
ggml_metal_init: skipping
kernel_cpy_bf16_f32                    (not
supported)
ggml_metal_init: skipping
kernel_cpy_bf16_bf16                  (not
supported)
ggml_metal_init: loaded
kernel_cpy_f32_q8_0
0x107884760 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_cpy_f32_q4_0
0x1267e6860 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_cpy_f32_q4_1
0x107884d10 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_cpy_f32_q5_0
0x107885430 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_cpy_f32_q5_1
0x1267e6e10 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_cpy_f32_iq4_nl
0x1267e7530 | th_max = 1024 | th_width =
32
```

```
ggml_metal_init: loaded
kernel_cpy_q4_0_f32
0x1266c4f10 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_cpy_q4_0_f16
0x107885870 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_cpy_q4_1_f32
0x107885f90 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_cpy_q4_1_f16
0x107886540 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_cpy_q5_0_f32
0x1267e7970 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_cpy_q5_0_f16
0x107886980 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
```

```
kernel_cpy_q5_1_f32
0x1078870a0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_cpy_q5_1_f16
0x1267e7f20 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_cpy_q8_0_f32
0x1078874e0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_cpy_q8_0_f16
0x1267e84d0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_concat
0x107887b40 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_sqr
0x107888190 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_sqrt
0x1078889a0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_sin
```

```
0x1078891b0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_cos
0x1078899c0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_neg
0x1267e8b20 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_reglu
0x1267e8f90 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_geglu
0x10672cc20 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_swiglu
0x10788a130 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_geglu_erf
0x10788a390 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_geglu_quick
0x107950360 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_sum_rows
0x10788adb0 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_mean
0x10788b010 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded kernel_argmax
0x10788b730 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_pool_2d_avg_f32
0x10788c410 | th_max = 1024 | th_width =
32
ggml_metal_init: loaded
kernel_pool_2d_max_f32
0x10788c970 | th_max = 1024 | th_width =
32
set_abort_callback: call
llama_context:        CPU  output buffer size =
0.49 MiB
create_memory: n_ctx = 4096 (padded)
llama_kv_cache_unified: layer   0: dev =
Metal
llama_kv_cache_unified: layer   1: dev = Metal
llama_kv_cache_unified: layer   2: dev =
Metal
llama_kv_cache_unified: layer   3: dev =
Metal
llama_kv_cache_unified: layer   4: dev =
Metal
llama_kv_cache_unified: layer   5: dev =
Metal
llama_kv_cache_unified: layer   6: dev =
Metal
llama_kv_cache_unified: layer   7: dev =
Metal
llama_kv_cache_unified: layer   8: dev =
Metal
llama_kv_cache_unified: layer   9: dev =
Metal
llama_kv_cache_unified: layer  10: dev =
Metal
llama_kv_cache_unified: layer  11: dev =
Metal
llama_kv_cache_unified: layer  12: dev =
Metal
llama_kv_cache_unified: layer  13: dev =
Metal
llama_kv_cache_unified: layer  14: dev =
Metal
llama_kv_cache_unified: layer  15: dev =
Metal
```

```
llama_kv_cache_unified: layer  16: dev =
Metal
llama_kv_cache_unified: layer  17: dev =
Metal
llama_kv_cache_unified: layer  18: dev =
Metal
llama_kv_cache_unified: layer  19: dev =
Metal
llama_kv_cache_unified: layer  20: dev =
Metal
llama_kv_cache_unified: layer  21: dev =
Metal
llama_kv_cache_unified: layer  22: dev =
Metal
llama_kv_cache_unified: layer  23: dev =
Metal
llama_kv_cache_unified: layer  24: dev =
Metal
llama_kv_cache_unified: layer  25: dev =
Metal
llama_kv_cache_unified: layer  26: dev =
Metal
llama_kv_cache_unified: layer  27: dev =
Metal
llama_kv_cache_unified: layer  28: dev =
Metal
llama_kv_cache_unified: layer  29: dev =
Metal
llama_kv_cache_unified: layer  30: dev =
Metal
llama_kv_cache_unified: layer  31: dev =
Metal
llama_kv_cache_unified:      Metal KV buffer
size =   512.00 MiB
llama_kv_cache_unified: size =  512.00 MiB
(  4096 cells,  32 layers,  1 seqs), K (f16):
256.00 MiB, V (f16):  256.00 MiB
llama_kv_cache_unified:
LLAMA_SET_ROWS=0, using old ggml_cpy()
method for backwards compatibility
llama_context: enumerating backends
llama_context: backend_ptrs.size() = 3
llama_context: max_nodes = 65536
llama_context: worst-case: n_tokens = 512,
n_seqs = 1, n_outputs = 0
graph_reserve: reserving a graph for ubatch
with n_tokens =  512, n_seqs =  1, n_outputs
=  512
graph_reserve: reserving a graph for ubatch
with n_tokens =    1, n_seqs =  1, n_outputs =
1
graph_reserve: reserving a graph for ubatch
with n_tokens =  512, n_seqs =  1, n_outputs
=  512
llama_context:      Metal compute buffer size
=   296.00 MiB
llama_context:        CPU compute buffer size
=    16.01 MiB
llama_context: graph nodes  = 1158
llama_context: graph splits = 2
Metal : EMBED_LIBRARY = 1 | CPU : NEON =
1 | ARM_FMA = 1 | FP16_VA = 1 | DOTPROD =
1 | LLAMAFILE = 1 | ACCELERATE = 1 |
REPACK = 1 |
Model metadata:
{'general.quantization_version': '2',
'tokenizer.chat_template': "{% set
loop_messages = messages %}{% for
message in loop_messages %}{% set content
= '<|start_header_id|>' + message['role'] +
'<|end_header_id|>\n\n'+
message['content'] | trim + '<|eot_id|>' %}{%
if loop×index0 == 0 %}{% set content =
bos_token + content %}{% endif %}
{{ content }}{% endfor %}{% if
add_generation_prompt %}{{ '<|
start_header_id|>assistant<|end_header_id|
>\n\n' }}{% endif %}",
'tokenizer.ggml.eos_token_id': '128001',
'tokenizer.ggml.bos_token_id': '128000',
'tokenizer.ggml.model': 'gpt2',
'llama.vocab_size': '128256',
'llama.attention.head_count_kv': '8',
'llama.context_length': '8192',
'llama.attention.head_count': '32',
'general.file_type': '15',
'llama.feed_forward_length': '14336',
'llama.rope.dimension_count': '128',
'llama.rope.freq_base': '500000.000000',
'llama.embedding_length': '4096',
'general.architecture': 'llama',
'llama.attention.layer_norm_rms_epsilon':
'0.000010', 'general.name': 'Meta-
Llama-3-8B-Instruct', 'llama.block_count':
'32'}
Available chat formats from metadata:
chat_template.default
Guessed chat format: llama-3
You: hello Jarvis
ggml_metal_graph_compute: command
```

```
buffer 1 failed with status 5
error: Insufficient Memory
(00000008:kIOGPUCommandBufferCallbac
kErrorOutOfMemory)
graph_compute:
ggml_backend_sched_graph_compute_asyn
c failed with error -1
process_ubatch: failed to compute graph,
compute status: -1
decode: removing KV cache entries for
seq_id = 0, pos = [0, +inf)
llama_decode: failed to decode, ret = -3
Traceback (most recent call last):
  File "/Users/andrebaker/llama.cpp/
jarvis_llama.py", line 24, in <module>
    response = llm(prompt, max_tokens=256)
  File "/Users/andrebaker/llama.cpp/llama-
voice-env/lib/python3.13/site-packages/
llama_cpp/llama.py", line 1904, in __call__
    return self.create_completion(
           ~~~~~~~~~~~~~~~~~~~~~~^
        prompt=prompt,
        ^^^^^^^^^^^^^^

    ...<23 lines>...
        logit_bias=logit_bias,
```

```
      ^^^^^^^^^^^^^^^^^^^^^^^^
    )
    ^
  File "/Users/andrebaker/llama.cpp/llama-
voice-env/lib/python3.13/site-packages/
llama_cpp/llama.py", line 1837, in
create_completion
    completion: Completion =
next(completion_or_chunks)  # type:
ignore

~~~~^^^^^^^^^^^^^^^^^^^^^^^^
  File "/Users/andrebaker/llama.cpp/llama-
voice-env/lib/python3.13/site-packages/
llama_cpp/llama.py", line 1322, in
_create_completion
    for token in self.generate(
                 ~~~~~~~~~~~~~^
        prompt_tokens,
        ^^^^^^^^^^^^^^
    ...<14 lines>...
        grammar=grammar,
        ^^^^^^^^^^^^^^^^
    ):
    ^
```

```
  File "/Users/andrebaker/llama.cpp/llama-
voice-env/lib/python3.13/site-packages/
llama_cpp/llama.py", line 914, in generate
    self.eval(tokens)
    ~~~~~~~~~^^^^^^^^
  File "/Users/andrebaker/llama.cpp/llama-
voice-env/lib/python3.13/site-packages/
llama_cpp/llama.py", line 648, in eval
    self._ctx.decode(self._batch)
    ~~~~~~~~~~~~~~~~^^^^^^^^^^^^^
  File "/Users/andrebaker/llama.cpp/llama-
voice-env/lib/python3.13/site-packages/
llama_cpp/_internals.py", line 321, in decode
    raise RuntimeError(f"llama_decode
returned {return_code}")
RuntimeError: llama_decode returned -3
ggml_metal_free: deallocating
ggml_metal_mem_pool_free: freeing
memory pool, num heaps = 0 (total = 0)
ggml_metal_mem_pool_free: freeing
memory pool, num heaps = 0 (total = 0)
ggml_metal_mem_pool_free: freeing
memory pool, num heaps = 0 (total = 0)
ggml_metal_mem_pool_free: freeing
memory pool, num heaps = 0 (total = 0)
```

```
ggml_metal_mem_pool_free: freeing
memory pool, num heaps = 0 (total = 0)
ggml_metal_mem_pool_free: freeing
memory pool, num heaps = 0 (total = 0)
ggml_metal_mem_pool_free: freeing
memory pool, num heaps = 0 (total = 0)
ggml_metal_mem_pool_free: freeing
memory pool, num heaps = 0 (total = 0)
🛑 LLaMA server stopped.
srv    operator(): operator(): cleaning up
before exit...

Saving session...ggml_metal_free:
deallocating

...copying shared history...
...saving history...truncating history files...
...completed.
Deleting expired sessions...none found.

[Process completed]
```