TEST VOICE PLC (CPU ONLY)  
  
(voicePLC) stanley@NUC12:~/coding/voicePLC$ python test\_voice\_plc.py

ggml\_cuda\_init: GGML\_CUDA\_FORCE\_MMQ: no

ggml\_cuda\_init: GGML\_CUDA\_FORCE\_CUBLAS: no

ggml\_cuda\_init: found 1 CUDA devices:

Device 0: NVIDIA GeForce RTX 4070 SUPER, compute capability 8.9, VMM: yes

llama\_load\_model\_from\_file: using device CUDA0 (NVIDIA GeForce RTX 4070 SUPER) - 8192 MiB free

llama\_model\_loader: loaded meta data with 24 key-value pairs and 291 tensors from models/mistral-7b-instruct-v0.2.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 = mistralai\_mistral-7b-instruct-v0.2

llama\_model\_loader: - kv 2: llama.context\_length u32 = 32768

llama\_model\_loader: - kv 3: llama.embedding\_length u32 = 4096

llama\_model\_loader: - kv 4: llama.block\_count u32 = 32

llama\_model\_loader: - kv 5: llama.feed\_forward\_length u32 = 14336

llama\_model\_loader: - kv 6: llama.rope.dimension\_count u32 = 128

llama\_model\_loader: - kv 7: llama.attention.head\_count u32 = 32

llama\_model\_loader: - kv 8: llama.attention.head\_count\_kv u32 = 8

llama\_model\_loader: - kv 9: llama.attention.layer\_norm\_rms\_epsilon f32 = 0.000010

llama\_model\_loader: - kv 10: llama.rope.freq\_base f32 = 1000000.000000

llama\_model\_loader: - kv 11: general.file\_type u32 = 15

llama\_model\_loader: - kv 12: tokenizer.ggml.model str = llama

llama\_model\_loader: - kv 13: tokenizer.ggml.tokens arr[str,32000] = ["<unk>", "<s>", "</s>", "<0x00>", "<...

llama\_model\_loader: - kv 14: tokenizer.ggml.scores arr[f32,32000] = [0.000000, 0.000000, 0.000000, 0.0000...

llama\_model\_loader: - kv 15: tokenizer.ggml.token\_type arr[i32,32000] = [2, 3, 3, 6, 6, 6, 6, 6, 6, 6, 6, 6, ...

llama\_model\_loader: - kv 16: tokenizer.ggml.bos\_token\_id u32 = 1

llama\_model\_loader: - kv 17: tokenizer.ggml.eos\_token\_id u32 = 2

llama\_model\_loader: - kv 18: tokenizer.ggml.unknown\_token\_id u32 = 0

llama\_model\_loader: - kv 19: tokenizer.ggml.padding\_token\_id u32 = 0

llama\_model\_loader: - kv 20: tokenizer.ggml.add\_bos\_token bool = true

llama\_model\_loader: - kv 21: tokenizer.ggml.add\_eos\_token bool = false

llama\_model\_loader: - kv 22: tokenizer.chat\_template str = {{ bos\_token }}{% for message in mess...

llama\_model\_loader: - kv 23: 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

llm\_load\_vocab: control token: 2 '</s>' is not marked as EOG

llm\_load\_vocab: control token: 1 '<s>' is not marked as EOG

llm\_load\_vocab: special\_eos\_id is not in special\_eog\_ids - the tokenizer config may be incorrect

llm\_load\_vocab: special tokens cache size = 3

llm\_load\_vocab: token to piece cache size = 0.1637 MB

llm\_load\_print\_meta: format = GGUF V3 (latest)

llm\_load\_print\_meta: arch = llama

llm\_load\_print\_meta: vocab type = SPM

llm\_load\_print\_meta: n\_vocab = 32000

llm\_load\_print\_meta: n\_merges = 0

llm\_load\_print\_meta: vocab\_only = 0

llm\_load\_print\_meta: n\_ctx\_train = 32768

llm\_load\_print\_meta: n\_embd = 4096

llm\_load\_print\_meta: n\_layer = 32

llm\_load\_print\_meta: n\_head = 32

llm\_load\_print\_meta: n\_head\_kv = 8

llm\_load\_print\_meta: n\_rot = 128

llm\_load\_print\_meta: n\_swa = 0

llm\_load\_print\_meta: n\_embd\_head\_k = 128

llm\_load\_print\_meta: n\_embd\_head\_v = 128

llm\_load\_print\_meta: n\_gqa = 4

llm\_load\_print\_meta: n\_embd\_k\_gqa = 1024

llm\_load\_print\_meta: n\_embd\_v\_gqa = 1024

llm\_load\_print\_meta: f\_norm\_eps = 0.0e+00

llm\_load\_print\_meta: f\_norm\_rms\_eps = 1.0e-05

llm\_load\_print\_meta: f\_clamp\_kqv = 0.0e+00

llm\_load\_print\_meta: f\_max\_alibi\_bias = 0.0e+00

llm\_load\_print\_meta: f\_logit\_scale = 0.0e+00

llm\_load\_print\_meta: n\_ff = 14336

llm\_load\_print\_meta: n\_expert = 0

llm\_load\_print\_meta: n\_expert\_used = 0

llm\_load\_print\_meta: causal attn = 1

llm\_load\_print\_meta: pooling type = 0

llm\_load\_print\_meta: rope type = 0

llm\_load\_print\_meta: rope scaling = linear

llm\_load\_print\_meta: freq\_base\_train = 1000000.0

llm\_load\_print\_meta: freq\_scale\_train = 1

llm\_load\_print\_meta: n\_ctx\_orig\_yarn = 32768

llm\_load\_print\_meta: rope\_finetuned = unknown

llm\_load\_print\_meta: ssm\_d\_conv = 0

llm\_load\_print\_meta: ssm\_d\_inner = 0

llm\_load\_print\_meta: ssm\_d\_state = 0

llm\_load\_print\_meta: ssm\_dt\_rank = 0

llm\_load\_print\_meta: ssm\_dt\_b\_c\_rms = 0

llm\_load\_print\_meta: model type = 7B

llm\_load\_print\_meta: model ftype = Q4\_K - Medium

llm\_load\_print\_meta: model params = 7.24 B

llm\_load\_print\_meta: model size = 4.07 GiB (4.83 BPW)

llm\_load\_print\_meta: general.name = mistralai\_mistral-7b-instruct-v0.2

llm\_load\_print\_meta: BOS token = 1 '<s>'

llm\_load\_print\_meta: EOS token = 2 '</s>'

llm\_load\_print\_meta: UNK token = 0 '<unk>'

llm\_load\_print\_meta: PAD token = 0 '<unk>'

llm\_load\_print\_meta: LF token = 13 '<0x0A>'

llm\_load\_print\_meta: EOG token = 2 '</s>'

llm\_load\_print\_meta: max token length = 48

llm\_load\_tensors: tensor 'token\_embd.weight' (q4\_K) (and 290 others) cannot be used with preferred buffer type CPU\_AARCH64, using CPU instead

llm\_load\_tensors: offloading 0 repeating layers to GPU

llm\_load\_tensors: offloaded 0/33 layers to GPU

llm\_load\_tensors: CPU\_Mapped model buffer size = 4165.37 MiB

.................................................................................................

llama\_new\_context\_with\_model: n\_seq\_max = 1

llama\_new\_context\_with\_model: n\_ctx = 512

llama\_new\_context\_with\_model: n\_ctx\_per\_seq = 512

llama\_new\_context\_with\_model: n\_batch = 512

llama\_new\_context\_with\_model: n\_ubatch = 512

llama\_new\_context\_with\_model: flash\_attn = 0

llama\_new\_context\_with\_model: freq\_base = 1000000.0

llama\_new\_context\_with\_model: freq\_scale = 1

llama\_new\_context\_with\_model: n\_ctx\_per\_seq (512) < n\_ctx\_train (32768) -- the full capacity of the model will not be utilized

llama\_kv\_cache\_init: CPU KV buffer size = 64.00 MiB

llama\_new\_context\_with\_model: KV self size = 64.00 MiB, K (f16): 32.00 MiB, V (f16): 32.00 MiB

llama\_new\_context\_with\_model: CPU output buffer size = 0.12 MiB

llama\_new\_context\_with\_model: CUDA0 compute buffer size = 181.04 MiB

llama\_new\_context\_with\_model: CUDA\_Host compute buffer size = 9.01 MiB

llama\_new\_context\_with\_model: graph nodes = 1030

llama\_new\_context\_with\_model: graph splits = 356 (with bs=512), 1 (with bs=1)

CUDA : ARCHS = 890 | USE\_GRAPHS = 1 | PEER\_MAX\_BATCH\_SIZE = 128 | CPU : SSE3 = 1 | SSSE3 = 1 | AVX = 1 | AVX\_VNNI = 1 | AVX2 = 1 | F16C = 1 | FMA = 1 | LLAMAFILE = 1 | OPENMP = 1 | AARCH64\_REPACK = 1 |

Model metadata: {'tokenizer.chat\_template': "{{ bos\_token }}{% for message in messages %}{% if (message['role'] == 'user') != (loop.index0 % 2 == 0) %}{{ raise\_exception('Conversation roles must alternate user/assistant/user/assistant/...') }}{% endif %}{% if message['role'] == 'user' %}{{ '[INST] ' + message['content'] + ' [/INST]' }}{% elif message['role'] == 'assistant' %}{{ message['content'] + eos\_token}}{% else %}{{ raise\_exception('Only user and assistant roles are supported!') }}{% endif %}{% endfor %}", 'tokenizer.ggml.add\_eos\_token': 'false', 'tokenizer.ggml.padding\_token\_id': '0', 'tokenizer.ggml.unknown\_token\_id': '0', 'tokenizer.ggml.eos\_token\_id': '2', 'general.architecture': 'llama', 'llama.rope.freq\_base': '1000000.000000', 'llama.context\_length': '32768', 'general.name': 'mistralai\_mistral-7b-instruct-v0.2', 'tokenizer.ggml.add\_bos\_token': 'true', 'llama.embedding\_length': '4096', 'llama.feed\_forward\_length': '14336', 'llama.attention.layer\_norm\_rms\_epsilon': '0.000010', 'llama.rope.dimension\_count': '128', 'tokenizer.ggml.bos\_token\_id': '1', 'llama.attention.head\_count': '32', 'llama.block\_count': '32', 'llama.attention.head\_count\_kv': '8', 'general.quantization\_version': '2', 'tokenizer.ggml.model': 'llama', 'general.file\_type': '15'}

Available chat formats from metadata: chat\_template.default

Guessed chat format: mistral-instruct

LOG (VoskAPI:ReadDataFiles():model.cc:213) Decoding params beam=10 max-active=3000 lattice-beam=2

LOG (VoskAPI:ReadDataFiles():model.cc:216) Silence phones 1:2:3:4:5:6:7:8:9:10

LOG (VoskAPI:RemoveOrphanNodes():nnet-nnet.cc:948) Removed 0 orphan nodes.

LOG (VoskAPI:RemoveOrphanComponents():nnet-nnet.cc:847) Removing 0 orphan components.

LOG (VoskAPI:ReadDataFiles():model.cc:248) Loading i-vector extractor from models/vosk-model-small-en-us-0.15/ivector/final.ie

LOG (VoskAPI:ComputeDerivedVars():ivector-extractor.cc:183) Computing derived variables for iVector extractor

LOG (VoskAPI:ComputeDerivedVars():ivector-extractor.cc:204) Done.

LOG (VoskAPI:ReadDataFiles():model.cc:282) Loading HCL and G from models/vosk-model-small-en-us-0.15/graph/HCLr.fst models/vosk-model-small-en-us-0.15/graph/Gr.fst

LOG (VoskAPI:ReadDataFiles():model.cc:308) Loading winfo models/vosk-model-small-en-us-0.15/graph/phones/word\_boundary.int

Starting voice recognition...

Speak commands or press Ctrl+C to exit

Try commands like:

- What is the temperature of sensor 1

- Turn actuator 1 on

- Get the pressure from sensor 2

Recognized: if meter readings from the sensor to

/home/stanley/miniconda3/envs/voicePLC/lib/python3.10/site-packages/llama\_cpp/llama.py:1237: RuntimeWarning: Detected duplicate leading "<s>" in prompt, this will likely reduce response quality, consider removing it...

warnings.warn(

llama\_perf\_context\_print: load time = 1278.74 ms

llama\_perf\_context\_print: prompt eval time = 0.00 ms / 159 tokens ( 0.00 ms per token, inf tokens per second)

llama\_perf\_context\_print: eval time = 0.00 ms / 24 runs ( 0.00 ms per token, inf tokens per second)

llama\_perf\_context\_print: total time = 5933.64 ms / 183 tokens

Response: Temperature reading from sensor\_1: 32.58°C

Playing WAVE '/tmp/tmpu8zllktl.wav' : Signed 16 bit Little Endian, Rate 22050 Hz, Mono

Recognized: what about from the sensor to

/home/stanley/miniconda3/envs/voicePLC/lib/python3.10/site-packages/llama\_cpp/llama.py:1237: RuntimeWarning: Detected duplicate leading "<s>" in prompt, this will likely reduce response quality, consider removing it...

warnings.warn(

Llama.generate: 147 prefix-match hit, remaining 10 prompt tokens to eval

llama\_perf\_context\_print: load time = 1278.74 ms

llama\_perf\_context\_print: prompt eval time = 0.00 ms / 10 tokens ( 0.00 ms per token, inf tokens per second)

llama\_perf\_context\_print: eval time = 0.00 ms / 27 runs ( 0.00 ms per token, inf tokens per second)

llama\_perf\_context\_print: total time = 5507.01 ms / 37 tokens

Response: Temperature reading from sensor\_1: 27.47°C

Playing WAVE '/tmp/tmpl6p8\_\_wf.wav' : Signed 16 bit Little Endian, Rate 22050 Hz, Mono

Recognized: what is the status of actuator to

/home/stanley/miniconda3/envs/voicePLC/lib/python3.10/site-packages/llama\_cpp/llama.py:1237: RuntimeWarning: Detected duplicate leading "<s>" in prompt, this will likely reduce response quality, consider removing it...

warnings.warn(

Llama.generate: 148 prefix-match hit, remaining 12 prompt tokens to eval

llama\_perf\_context\_print: load time = 1278.74 ms

llama\_perf\_context\_print: prompt eval time = 0.00 ms / 12 tokens ( 0.00 ms per token, inf tokens per second)

llama\_perf\_context\_print: eval time = 0.00 ms / 29 runs ( 0.00 ms per token, inf tokens per second)

llama\_perf\_context\_print: total time = 5937.43 ms / 41 tokens

Response: Set actuator\_1 to OFF

Playing WAVE '/tmp/tmppnukzgkc.wav' : Signed 16 bit Little Endian, Rate 22050 Hz, Mono

Recognized: he said actuator one to on

/home/stanley/miniconda3/envs/voicePLC/lib/python3.10/site-packages/llama\_cpp/llama.py:1237: RuntimeWarning: Detected duplicate leading "<s>" in prompt, this will likely reduce response quality, consider removing it...

warnings.warn(

Llama.generate: 147 prefix-match hit, remaining 12 prompt tokens to eval

llama\_perf\_context\_print: load time = 1278.74 ms

llama\_perf\_context\_print: prompt eval time = 0.00 ms / 12 tokens ( 0.00 ms per token, inf tokens per second)

llama\_perf\_context\_print: eval time = 0.00 ms / 31 runs ( 0.00 ms per token, inf tokens per second)

llama\_perf\_context\_print: total time = 6366.19 ms / 43 tokens

Response: Set actuator\_1 to ON

Playing WAVE '/tmp/tmp52fjgb7e.wav' : Signed 16 bit Little Endian, Rate 22050 Hz, Mono