# crosscheck_llama-2-7b

April 18, 2024

[ ]:
```
import time
print("_____________INITIALIZING MODEL_____________")
time.sleep(1)
from llama_cpp import Llama
import json

def execute_localLLM(qas):
    normalized_responses = []
    i = 1
    for pair in qas:
        # generate prompt
        message = [{
            "role": "user",
            "content": f"Determine if this question and answer pair is fact or␣
 ↪humor. The question is: {pair['question']} The answer is: {pair['answer']}"
        }]

        print("\n_____________MODEL GENERATING RESPONSE_____________")
        # Generate a response
        response = llm.
 ↪create_chat_completion(messages=message)["choices"][0]["message"]["content"]
        print("_________________RESPONSE GENERATED_________________\n")

        # Print the response
        print("Response Number ", i)
        print(response)

        if "humor" in response.lower():
            normalized_responses.append(True)
        else:
            normalized_responses.append(False)
        i += 1
    return normalized_responses

def LLM_accuracy(LLM_responses, true_values):
    num_correct = 0
    for i in range(len(LLM_responses)):
```

```
        if LLM_responses[i] == true_values[i]['humor']:
            num_correct += 1
    return num_correct, num_correct / len(LLM_responses)


model_path = "C:/Users/edgar/Software/LLMs/text-generation-webui-main/models/
 ↪llama-2-7b-chat.Q5_K_M.gguf"

llm = Llama(model_path=model_path)
print("__________INITIALIZATION COMPLETE__________\n")

with open('../datastore/Questions_and_Answers.json', 'r') as f:
    qas = json.load(f)
qas = qas[:25]

normalized_responses = execute_localLLM(qas)
print("\n_________________Local LLM execution complete!_________________")

print("\nResponses from Local LLM normalized to true and false answers:\n",␣
 ↪normalized_responses)

num_correct, accuracy = LLM_accuracy(normalized_responses, qas)

print(f"\nLocal LLM model llama-2-7b identified {num_correct} out of {len(qas)}␣
 ↪question-answer pairs correctly. Accuracy = {accuracy*100:.2f}%")
```

```
______________INITIALIZING MODEL______________
llama_model_loader: loaded meta data with 19 key-value pairs and 291 tensors
from C:/Users/edgar/Software/LLMs/text-generation-webui-
main/models/llama-2-7b-chat.Q5_K_M.gguf (version GGUF V2)
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
= LLaMA v2
llama_model_loader: - kv   2:                      llama.context_length u32
= 4096
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
= 11008
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
= 32
llama_model_loader: - kv   9:     llama.attention.layer_norm_rms_epsilon f32
= 0.000001
llama_model_loader: - kv  10:                          general.file_type u32
= 17
llama_model_loader: - kv  11:                       tokenizer.ggml.model str
= llama
llama_model_loader: - kv  12:                      tokenizer.ggml.tokens
arr[str,32000]   = ["<unk>", "<s>", "</s>", "<0x00>", "<…
llama_model_loader: - kv  13:                      tokenizer.ggml.scores
arr[f32,32000]   = [0.000000, 0.000000, 0.000000, 0.0000…
llama_model_loader: - kv  14:                  tokenizer.ggml.token_type
arr[i32,32000]   = [2, 3, 3, 6, 6, 6, 6, 6, 6, 6, 6, 6, …
llama_model_loader: - kv  15:                tokenizer.ggml.bos_token_id u32
= 1
llama_model_loader: - kv  16:                tokenizer.ggml.eos_token_id u32
= 2
llama_model_loader: - kv  17:            tokenizer.ggml.unknown_token_id u32
= 0
llama_model_loader: - kv  18:               general.quantization_version u32
= 2
llama_model_loader: - type  f32:   65 tensors
llama_model_loader: - type q5_K:  193 tensors
llama_model_loader: - type q6_K:   33 tensors
llm_load_vocab: special tokens definition check successful ( 259/32000 ).
llm_load_print_meta: format           = GGUF V2
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: n_ctx_train      = 4096
llm_load_print_meta: n_embd           = 4096
llm_load_print_meta: n_head           = 32
llm_load_print_meta: n_head_kv        = 32
llm_load_print_meta: n_layer          = 32
llm_load_print_meta: n_rot            = 128
llm_load_print_meta: n_embd_head_k    = 128
llm_load_print_meta: n_embd_head_v    = 128
llm_load_print_meta: n_gqa            = 1
llm_load_print_meta: n_embd_k_gqa     = 4096
llm_load_print_meta: n_embd_v_gqa     = 4096
llm_load_print_meta: f_norm_eps       = 0.0e+00
llm_load_print_meta: f_norm_rms_eps   = 1.0e-06
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             = 11008
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  = 10000.0
llm_load_print_meta: freq_scale_train = 1
llm_load_print_meta: n_yarn_orig_ctx  = 4096
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: model type       = 7B
llm_load_print_meta: model ftype      = Q5_K - Medium
llm_load_print_meta: model params     = 6.74 B
llm_load_print_meta: model size       = 4.45 GiB (5.68 BPW)
llm_load_print_meta: general.name     = LLaMA v2
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: LF token         = 13 '<0x0A>'
llm_load_tensors: ggml ctx size =    0.11 MiB
llm_load_tensors:        CPU buffer size =  4560.87 MiB
…
…
llama_new_context_with_model: n_ctx      = 512
llama_new_context_with_model: n_batch    = 512
llama_new_context_with_model: n_ubatch   = 512
llama_new_context_with_model: freq_base  = 10000.0
llama_new_context_with_model: freq_scale = 1
llama_kv_cache_init:        CPU KV buffer size =   256.00 MiB
llama_new_context_with_model: KV self size  =  256.00 MiB, K (f16):  128.00 MiB,
V (f16):  128.00 MiB
llama_new_context_with_model:        CPU  output buffer size =     0.12 MiB
llama_new_context_with_model:        CPU compute buffer size =    70.50 MiB
llama_new_context_with_model: graph nodes  = 1030
llama_new_context_with_model: graph splits = 1
AVX = 1 | AVX_VNNI = 0 | AVX2 = 1 | AVX512 = 0 | AVX512_VBMI = 0 | AVX512_VNNI =
0 | FMA = 1 | NEON = 0 | ARM_FMA = 0 | F16C = 1 | FP16_VA = 0 | WASM_SIMD = 0 |
BLAS = 0 | SSE3 = 1 | SSSE3 = 0 | VSX = 0 | MATMUL_INT8 = 0 |
Model metadata: {'general.name': 'LLaMA v2', 'general.architecture': 'llama',
'llama.context_length': '4096', 'llama.rope.dimension_count': '128',
'llama.embedding_length': '4096', 'llama.block_count': '32',
'llama.feed_forward_length': '11008', 'llama.attention.head_count': '32',
'tokenizer.ggml.eos_token_id': '2', 'general.file_type': '17',
```

```
'llama.attention.head_count_kv': '32', 'llama.attention.layer_norm_rms_epsilon':
'0.000001', 'tokenizer.ggml.model': 'llama', 'general.quantization_version':
'2', 'tokenizer.ggml.bos_token_id': '1', 'tokenizer.ggml.unknown_token_id': '0'}
Using fallback chat format: None

__________INITIALIZATION COMPLETE__________


______________MODEL GENERATING RESPONSE______________

llama_print_timings:        load time =    2587.06 ms
llama_print_timings:      sample time =      11.55 ms /    86 runs   (    0.13
ms per token,  7443.31 tokens per second)
llama_print_timings: prompt eval time =    2586.83 ms /    50 tokens (   51.74
ms per token,    19.33 tokens per second)
llama_print_timings:        eval time =    8292.66 ms /    85 runs   (   97.56
ms per token,    10.25 tokens per second)
llama_print_timings:       total time =   11032.87 ms /   135 tokens

_________________RESPONSE GENERATED_________________

Response Number  1
  This is a humorous answer. The question is serious, asking for a commonality
between two seemingly unrelated things (gingers and extinct dinosaurs), but the
answer is a playful and tongue-in-cheek response that implies there is nothing
significant they have in common. It's a lighthearted and humorous take on the
question, rather than a factual or serious response.

______________MODEL GENERATING RESPONSE______________

Llama.generate: prefix-match hit

llama_print_timings:        load time =    2587.06 ms
llama_print_timings:      sample time =      20.95 ms /   153 runs   (    0.14
ms per token,  7302.41 tokens per second)
llama_print_timings: prompt eval time =    1543.36 ms /    31 tokens (   49.79
ms per token,    20.09 tokens per second)
llama_print_timings:        eval time =   14935.11 ms /   152 runs   (   98.26
ms per token,    10.18 tokens per second)
llama_print_timings:       total time =   16770.26 ms /   183 tokens
Llama.generate: prefix-match hit

_________________RESPONSE GENERATED_________________

Response Number  2
  This question and answer pair is likely meant to be humorous rather than
factual.
The region of Thuringia does not have a clear geographical location that would
align it with either the Germanic or Slavic territories in the Middle Ages. In
```

```
fact, Thuringia has a complex history that spans multiple cultures and time
periods, including the Frankish Empire, the Holy Roman Empire, and the German
Confederation.
The answer provided is likely intended to be humorous because it oversimplifies
the historical context of Thuringia and presents an absurd location for the
region in the Middle Ages. It is not a serious or accurate description of
Thuringia's geographical location during that time period.

_____________MODEL GENERATING RESPONSE_____________


llama_print_timings:        load time =    2587.06 ms
llama_print_timings:      sample time =      10.80 ms /    72 runs   (    0.15
ms per token,  6668.52 tokens per second)
llama_print_timings: prompt eval time =    1655.36 ms /    30 tokens (   55.18
ms per token,    18.12 tokens per second)
llama_print_timings:        eval time =    6965.05 ms /    71 runs   (   98.10
ms per token,    10.19 tokens per second)
llama_print_timings:       total time =    8760.02 ms /   101 tokens
Llama.generate: prefix-match hit

_________________RESPONSE GENERATED_________________

Response Number  3
  This question and answer pair is intended to be humorous, rather than factual.
The joke relies on a play on words, with the answer "There are twenty of them"
being a pun on the number of years in the age range mentioned in the question.
It is not meant to be taken seriously or as a factual statement.

_____________MODEL GENERATING RESPONSE_____________


llama_print_timings:        load time =    2587.06 ms
llama_print_timings:      sample time =      10.48 ms /    69 runs   (    0.15
ms per token,  6582.71 tokens per second)
llama_print_timings: prompt eval time =    1478.68 ms /    26 tokens (   56.87
ms per token,    17.58 tokens per second)
llama_print_timings:        eval time =    6676.54 ms /    68 runs   (   98.18
ms per token,    10.18 tokens per second)
llama_print_timings:       total time =    8293.21 ms /    94 tokens
Llama.generate: prefix-match hit

_________________RESPONSE GENERATED_________________

Response Number  4
  This is a humorous answer. Richard Feynman passed away on February 15, 1988,
so it would not have been possible for him to visit Tuva after he died. The
answer is a playful jab at the idea that Feynman could somehow travel through
time or defy death.
```

```
______________MODEL GENERATING RESPONSE_____________

llama_print_timings:        load time =    2587.06 ms
llama_print_timings:      sample time =      19.38 ms /   127 runs   (    0.15
ms per token,  6552.47 tokens per second)
llama_print_timings: prompt eval time =    1757.60 ms /    31 tokens (   56.70
ms per token,    17.64 tokens per second)
llama_print_timings:        eval time =   12442.88 ms /   126 runs   (   98.75
ms per token,    10.13 tokens per second)
llama_print_timings:       total time =   14456.20 ms /   157 tokens
Llama.generate: prefix-match hit

_________________RESPONSE GENERATED_________________

Response Number  5
  This question and answer pair is likely humor.
The question asks about the main function of Lundenwic by the late 7th century,
but the answer given is "a major port." However, there is no historical evidence
to suggest that Lundenwic was ever a major port. In fact, Lundenwic is believed
to have been a fictional city created by J.R.R. Tolkien for his Middle-earth
stories.
Therefore, the answer provided is likely intended as a humorous or satirical
response to the question, rather than a serious historical answer.

______________MODEL GENERATING RESPONSE_____________

llama_print_timings:        load time =    2587.06 ms
llama_print_timings:      sample time =      35.73 ms /   236 runs   (    0.15
ms per token,  6604.17 tokens per second)
llama_print_timings: prompt eval time =    1754.54 ms /    30 tokens (   58.48
ms per token,    17.10 tokens per second)
llama_print_timings:        eval time =   23685.30 ms /   235 runs   (  100.79
ms per token,     9.92 tokens per second)
llama_print_timings:       total time =   25932.61 ms /   265 tokens
Llama.generate: prefix-match hit

_________________RESPONSE GENERATED_________________

Response Number  6
  This question and answer pair is likely a joke or a play on words, rather than
a factual statement.
The idea that the British people "locked in national solidarity" during World
War II is an oxymoron, as it implies that the entire population of Britain was
somehow magically united in a state of complete agreement and unity, which is
highly unlikely. In reality, the British people were likely as divided as any
other population during times of war, with different opinions and beliefs on how
to handle the conflict.
Furthermore, the answer given is a bit too simplistic and doesn't take into
```

```
account the complexities of British society during WWII. The war had a profound
impact on Britain, both economically and socially, and it is unlikely that the
population was able to come together in a single, unified front.
Therefore, based on the language used and the lack of historical evidence to
support the claim, it is more likely that this question and answer pair is
intended as a humorous or satirical take on the idea of national solidarity
during wartime.

_____________MODEL GENERATING RESPONSE_____________


llama_print_timings:        load time =    2587.06 ms
llama_print_timings:      sample time =      15.53 ms /   102 runs   (    0.15
ms per token,  6568.78 tokens per second)
llama_print_timings: prompt eval time =    1353.37 ms /    24 tokens (   56.39
ms per token,    17.73 tokens per second)
llama_print_timings:        eval time =   10193.73 ms /   101 runs   (  100.93
ms per token,     9.91 tokens per second)
llama_print_timings:       total time =   11759.62 ms /   125 tokens
Llama.generate: prefix-match hit

_________________RESPONSE GENERATED_________________

Response Number  7
  This question and answer pair is likely meant to be humorous rather than
factual. The punchline of the answer, "Not being last," is a play on words that
pokes fun at the idea of comparing achievements in the Special Olympics, which
is an organization that provides opportunities for people with intellectual
disabilities to participate in sports and other activities. It's not meant to be
taken seriously or as a factual statement, but rather as a lighthearted joke.

_____________MODEL GENERATING RESPONSE_____________


llama_print_timings:        load time =    2587.06 ms
llama_print_timings:      sample time =      19.42 ms /   125 runs   (    0.16
ms per token,  6436.66 tokens per second)
llama_print_timings: prompt eval time =    1968.01 ms /    35 tokens (   56.23
ms per token,    17.78 tokens per second)
llama_print_timings:        eval time =   12402.53 ms /   124 runs   (  100.02
ms per token,    10.00 tokens per second)
llama_print_timings:       total time =   14629.49 ms /   159 tokens
Llama.generate: prefix-match hit

_________________RESPONSE GENERATED_________________

Response Number  8
  This is a humorous response. The initial question seems to be serious, asking
if the listener knows that a serial number is printed on every condom. However,
the punchline "I guess you haven't rolled it down far enough" is a playful and
```

unexpected twist that suggests the speaker is joking or teasing. It implies that the serial number might not actually be on the condom after all, perhaps due to rolling or other actions taken during sexual activity. The humor in this response likely comes from the unexpected turn and the idea of a serial number being hidden in an unexpected location.

```
______________MODEL GENERATING RESPONSE______________

llama_print_timings:        load time =    2587.06 ms
llama_print_timings:      sample time =      15.67 ms /   102 runs   (    0.15
ms per token,  6508.84 tokens per second)
llama_print_timings: prompt eval time =    1607.46 ms /    29 tokens (   55.43
ms per token,    18.04 tokens per second)
llama_print_timings:        eval time =   10106.71 ms /   101 runs   (  100.07
ms per token,     9.99 tokens per second)
llama_print_timings:       total time =   11924.29 ms /   130 tokens
Llama.generate: prefix-match hit

_________________RESPONSE GENERATED_________________

Response Number  9
  This is a humorous answer. The question is asking about something that is
impossible in the real world, as Jedi are fictional characters from the Star
Wars franchise and do not actually exist or use email. The answer is a playful
explanation for why someone cannot email a photo to a Jedi, using a logical but
absurd reason (attachments are not allowed) to add to the humor. Therefore, this
question and answer pair is intended to be humorous rather than factual.

______________MODEL GENERATING RESPONSE______________

llama_print_timings:        load time =    2587.06 ms
llama_print_timings:      sample time =      23.98 ms /   159 runs   (    0.15
ms per token,  6630.25 tokens per second)
llama_print_timings: prompt eval time =    1367.36 ms /    24 tokens (   56.97
ms per token,    17.55 tokens per second)
llama_print_timings:        eval time =   15657.87 ms /   158 runs   (   99.10
ms per token,    10.09 tokens per second)
llama_print_timings:       total time =   17351.48 ms /   182 tokens
Llama.generate: prefix-match hit

_________________RESPONSE GENERATED_________________

Response Number  10
  This question and answer pair is likely meant as a joke or a play on words,
rather than a serious factual statement.
The Polish nobility were known for their elaborate coats of arms, which were
symbols of their status and lineage. However, the idea that "Coats of arms were
very important to the Polish nobility" is an obvious exaggeration, as it implies
```

```
that they were the only thing that mattered to them, which is not true.
The answer is meant to be humorous because it takes a stereotype or cliché about
the Polish nobility (their love of coats of arms) and amplifies it to an absurd
degree, making it clear that it's not meant to be taken literally.

______________MODEL GENERATING RESPONSE______________

llama_print_timings:        load time =    2587.06 ms
llama_print_timings:      sample time =      16.34 ms /   107 runs   (    0.15
ms per token,  6547.15 tokens per second)
llama_print_timings: prompt eval time =    3252.62 ms /    60 tokens (   54.21
ms per token,    18.45 tokens per second)
llama_print_timings:        eval time =   10427.88 ms /   106 runs   (   98.38
ms per token,    10.17 tokens per second)
llama_print_timings:       total time =   13897.35 ms /   166 tokens
Llama.generate: prefix-match hit

_________________RESPONSE GENERATED_________________

Response Number  11
  This question and answer pair is humor. The statement "the widespread
irregularities make it impossible to know for certain that the outcome reflected
the will of the voters" is an obvious exaggeration and a play on words, implying
that there were so many irregularities in the Ohio vote that it's impossible to
determine the true outcome. It's not a serious or factual statement, but rather
a humorous take on the idea that election results can be unreliable or biased.

______________MODEL GENERATING RESPONSE______________

llama_print_timings:        load time =    2587.06 ms
llama_print_timings:      sample time =      15.14 ms /    98 runs   (    0.15
ms per token,  6473.77 tokens per second)
llama_print_timings: prompt eval time =    1374.65 ms /    24 tokens (   57.28
ms per token,    17.46 tokens per second)
llama_print_timings:        eval time =    9654.52 ms /    97 runs   (   99.53
ms per token,    10.05 tokens per second)
llama_print_timings:       total time =   11233.11 ms /   121 tokens
Llama.generate: prefix-match hit

_________________RESPONSE GENERATED_________________

Response Number  12
  This question and answer pair is likely meant to be humorous rather than
factual. The term "fabourite" is a play on the word "favorite," which is a
common spelling mistake, and the punchline "Unknown" is a tongue-in-cheek
response that pokes fun at the idea of asking about a transgender person's
favorite Pokémon. It is not meant to be taken seriously or as a factual answer.
```

```
______________MODEL GENERATING RESPONSE_____________

llama_print_timings:        load time =    2587.06 ms
llama_print_timings:      sample time =      24.02 ms /   147 runs   (    0.16
ms per token,  6119.65 tokens per second)
llama_print_timings: prompt eval time =    1406.32 ms /    25 tokens (   56.25
ms per token,    17.78 tokens per second)
llama_print_timings:        eval time =   16563.03 ms /   146 runs   (  113.45
ms per token,     8.81 tokens per second)
llama_print_timings:       total time =   18299.50 ms /   171 tokens
Llama.generate: prefix-match hit

_________________RESPONSE GENERATED_________________

Response Number  13
  This question and answer pair is likely a joke or a play on words, rather than
a factual question and answer.
The movie "Destry Rides Again" was actually released in 1939, as stated in the
answer. However, the question is phrased in a way that suggests it might be
asking for the year of release of a different movie or media property, such as
"The Lone Ranger" or "Batman," which are both popular franchises that were also
released in the 1930s.
Therefore, the answer of 1939 is likely intended to be humorous or misleading,
rather than a straightforward factual answer.

______________MODEL GENERATING RESPONSE_____________

llama_print_timings:        load time =    2587.06 ms
llama_print_timings:      sample time =      20.93 ms /   134 runs   (    0.16
ms per token,  6402.29 tokens per second)
llama_print_timings: prompt eval time =    1066.60 ms /    18 tokens (   59.26
ms per token,    16.88 tokens per second)
llama_print_timings:        eval time =   13561.32 ms /   133 runs   (  101.96
ms per token,     9.81 tokens per second)
llama_print_timings:       total time =   14914.26 ms /   151 tokens
Llama.generate: prefix-match hit

_________________RESPONSE GENERATED_________________

Response Number  14
  This question and answer pair is likely meant to be humorous rather than
factual. Here's why:
The question asks about a business that has closed in Macau, but the answer
given is "Casinos." While it is true that casinos are a major industry in Macau,
it is unlikely that all casinos in the region have closed. In fact, Macau is
known for its vibrant casino industry and is often referred to as the "Las Vegas
of Asia."
Therefore, the answer provided is likely intended to be humorous or satirical,
```

```
rather than a serious response to the question.

______________MODEL GENERATING RESPONSE______________


llama_print_timings:        load time =    2587.06 ms
llama_print_timings:      sample time =      14.81 ms /    95 runs   (    0.16
ms per token,  6413.72 tokens per second)
llama_print_timings: prompt eval time =    1458.81 ms /    26 tokens (   56.11
ms per token,    17.82 tokens per second)
llama_print_timings:        eval time =    9772.58 ms /    94 runs   (  103.96
ms per token,     9.62 tokens per second)
llama_print_timings:       total time =   11432.89 ms /   120 tokens
Llama.generate: prefix-match hit

_________________RESPONSE GENERATED_________________

Response Number  15
  This question and answer pair is intended to be humorous. The punchline
"Names!" is a play on words, as it can be interpreted in multiple ways,
including the idea that the person with no arms and an eye patch has a unique
name, but also implying that the question itself is absurd and doesn't deserve a
serious answer. It is not a factual or informative response, but rather a
lighthearted joke.

______________MODEL GENERATING RESPONSE______________


llama_print_timings:        load time =    2587.06 ms
llama_print_timings:      sample time =      15.51 ms /    98 runs   (    0.16
ms per token,  6317.69 tokens per second)
llama_print_timings: prompt eval time =    2130.16 ms /    38 tokens (   56.06
ms per token,    17.84 tokens per second)
llama_print_timings:        eval time =    9942.28 ms /    97 runs   (  102.50
ms per token,     9.76 tokens per second)
llama_print_timings:       total time =   12278.33 ms /   135 tokens
Llama.generate: prefix-match hit

_________________RESPONSE GENERATED_________________

Response Number  16
  This is a humorous answer to a question about the Jewish version of the board
game Monopoly. The punchline, "Because the banker starts with all the money and
never gives it away," is a play on stereotypes about Jews being cheap or greedy,
which is not a serious or factual explanation for why the game might be more
challenging. Therefore, this question and answer pair is intended to be humorous
rather than factual.

______________MODEL GENERATING RESPONSE______________
```

```
llama_print_timings:        load time =    2587.06 ms
llama_print_timings:      sample time =      12.44 ms /    82 runs   (    0.15
ms per token,  6593.23 tokens per second)
llama_print_timings: prompt eval time =    1545.83 ms /    27 tokens (   57.25
ms per token,    17.47 tokens per second)
llama_print_timings:        eval time =    8007.76 ms /    81 runs   (   98.86
ms per token,    10.12 tokens per second)
llama_print_timings:       total time =    9723.71 ms /   108 tokens
Llama.generate: prefix-match hit
```

```
_________________RESPONSE GENERATED_________________
```

```
Response Number  17
  This is a humorous answer. It is not a serious or factual response, but rather
a play on words. The punchline "You already told her" implies that the speaker
has previously said something to the girl with a black eye, but it is not a real
or helpful comment. It is meant to be amusing and lighthearted, rather than
providing actual advice or insight.
```

```
______________MODEL GENERATING RESPONSE______________
```

```
llama_print_timings:        load time =    2587.06 ms
llama_print_timings:      sample time =      11.48 ms /    77 runs   (    0.15
ms per token,  6704.98 tokens per second)
llama_print_timings: prompt eval time =    1981.03 ms /    37 tokens (   53.54
ms per token,    18.68 tokens per second)
llama_print_timings:        eval time =    7447.83 ms /    76 runs   (   98.00
ms per token,    10.20 tokens per second)
llama_print_timings:       total time =    9584.24 ms /   113 tokens
Llama.generate: prefix-match hit
```

```
_________________RESPONSE GENERATED_________________
```

```
Response Number  18
  This is a humorous answer. The question is asking for something serious and
meaningful that President Trump might have said to former President Obama, but
the answer given is a play on words and a reference to a popular TV show,
"Orange is the New Black." It's not a factual or serious response, but rather a
joke or meme.
```

```
______________MODEL GENERATING RESPONSE______________
```

```
llama_print_timings:        load time =    2587.06 ms
llama_print_timings:      sample time =       9.62 ms /    64 runs   (    0.15
ms per token,  6654.19 tokens per second)
llama_print_timings: prompt eval time =    1339.65 ms /    24 tokens (   55.82
ms per token,    17.92 tokens per second)
```

```
llama_print_timings:        eval time =    6137.08 ms /    63 runs   (   97.41
ms per token,    10.27 tokens per second)
llama_print_timings:       total time =    7605.37 ms /    87 tokens
Llama.generate: prefix-match hit
```

_____RESPONSE GENERATED_____

Response Number 19

This is a humorous question and answer pair. The pun on "calc" (short for calculate) being used as a play on words with "call" is a clever and amusing joke. It is not a factual or serious question, but rather a lighthearted and humorous exchange.

_____MODEL GENERATING RESPONSE_____

```
llama_print_timings:        load time =    2587.06 ms
llama_print_timings:      sample time =      39.46 ms /   260 runs   (    0.15
ms per token,  6588.62 tokens per second)
llama_print_timings: prompt eval time =    1170.34 ms /    20 tokens (   58.52
ms per token,    17.09 tokens per second)
llama_print_timings:        eval time =   25818.94 ms /   259 runs   (   99.69
ms per token,    10.03 tokens per second)
llama_print_timings:       total time =   27548.63 ms /   279 tokens
Llama.generate: prefix-match hit
```

_____RESPONSE GENERATED_____

Response Number 20

This question and answer pair is likely meant to be humorous rather than factual. Here's why:

1. The question is phrased in a way that implies there might be something unusual or surprising about Roman law being older than Catholic law, which suggests it's not a straightforward or obvious answer.
2. The answer provided, "Roman law," is a well-known and established legal system that has been around for thousands of years, long before the establishment of the Catholic Church. It's unlikely that anyone would be unaware of this fact, which makes it seem like the question is being asked in a tongue-in-cheek manner.
3. The humor may also come from the idea that the question is asking about the relative ages of two legal systems that are often seen as having a complex and sometimes contentious relationship. By implying that one system (Catholic law) might be older than the other (Roman law), the question is poking fun at the idea that there could be any real competition or hierarchy between these two legal traditions.

Overall, while it's possible that the question and answer pair could be meant to be taken literally, it seems more likely that they are intended as a humorous exchange.

```
______________MODEL GENERATING RESPONSE______________

llama_print_timings:        load time =    2587.06 ms
llama_print_timings:      sample time =      18.78 ms /   125 runs   (    0.15
ms per token,  6655.31 tokens per second)
llama_print_timings: prompt eval time =    2239.95 ms /    41 tokens (   54.63
ms per token,    18.30 tokens per second)
llama_print_timings:        eval time =   12427.62 ms /   124 runs   (  100.22
ms per token,     9.98 tokens per second)
llama_print_timings:       total time =   14929.16 ms /   165 tokens
Llama.generate: prefix-match hit

_________________RESPONSE GENERATED_________________

Response Number  21
  This question and answer pair is likely meant to be humorous rather than
factual.
The term "pastiche personality" is not a recognized psychological or
sociological term, and it is unlikely that there is a specific classification
for individuals who adopt social perceptions of themselves rather than their
true selves. The term seems to be a playful invention meant to poke fun at the
idea of people presenting a false or artificial persona to the world.
Therefore, the answer "The pastiche personality" is likely intended as a
humorous response rather than a serious classification.

______________MODEL GENERATING RESPONSE______________

llama_print_timings:        load time =    2587.06 ms
llama_print_timings:      sample time =      15.14 ms /    96 runs   (    0.16
ms per token,  6339.14 tokens per second)
llama_print_timings: prompt eval time =    1584.90 ms /    29 tokens (   54.65
ms per token,    18.30 tokens per second)
llama_print_timings:        eval time =   10175.09 ms /    95 runs   (  107.11
ms per token,     9.34 tokens per second)
llama_print_timings:       total time =   11974.10 ms /   124 tokens
Llama.generate: prefix-match hit

_________________RESPONSE GENERATED_________________

Response Number  22
  This question and answer pair is likely humor. The question sets up an
expectation that Buddhism would take a particular stance on the spectrum of
pessimistic vs optimistic, but the answer provides a unexpected and seemingly
contradictory response by saying "realistic." This unexpected twist suggests
that the answer may not be meant to be taken literally, but rather as a humorous
jab at the idea that Buddhism could be reduced to a simple binary opposition.

______________MODEL GENERATING RESPONSE______________
```

```
llama_print_timings:        load time =    2587.06 ms
llama_print_timings:      sample time =      34.51 ms /   221 runs   (    0.16
ms per token,  6404.13 tokens per second)
llama_print_timings: prompt eval time =    1279.81 ms /    23 tokens (   55.64
ms per token,    17.97 tokens per second)
llama_print_timings:        eval time =   23653.08 ms /   220 runs   (  107.51
ms per token,     9.30 tokens per second)
llama_print_timings:       total time =   25428.03 ms /   243 tokens
Llama.generate: prefix-match hit

_________________RESPONSE GENERATED_________________

Response Number  23
  This question and answer pair is likely meant as a joke or a play on words,
rather than a serious statement of fact.
The term "five precepts" typically refers to the basic ethical principles in
Buddhism, which are:
1. Refrain from killing or harming living beings
2. Refrain from stealing or taking what is not given
3. Refrain from sexual misconduct
4. Refrain from lying, gossiping, or speaking harshly to others
5. Refrain from intoxicating substances that cloud the mind

The question "Lay people tend to live by the five what?" sets up an expectation
that the answer will be one of these Buddhist precepts, but the punchline
"precepts" is a pun on the word "principles" or "guidelines," which is not a
serious or accurate interpretation of the term "five precepts." Therefore, this
question and answer pair is likely meant as humor rather than a factual
statement.

_____________MODEL GENERATING RESPONSE_____________


llama_print_timings:        load time =    2587.06 ms
llama_print_timings:      sample time =      38.88 ms /   257 runs   (    0.15
ms per token,  6610.08 tokens per second)
llama_print_timings: prompt eval time =    1617.97 ms /    29 tokens (   55.79
ms per token,    17.92 tokens per second)
llama_print_timings:        eval time =   25401.51 ms /   256 runs   (   99.22
ms per token,    10.08 tokens per second)
llama_print_timings:       total time =   27589.08 ms /   285 tokens
Llama.generate: prefix-match hit

_________________RESPONSE GENERATED_________________

Response Number  24
  This question and answer pair is likely meant as humor rather than a factual
statement. Here's why:
```

```
1. The question asks about a specific country, Afghanistan, but the answer given
is a species that is not found in Afghanistan or Asia. Siberian cranes are
actually found in Russia, China, and other parts of Asia, not in Afghanistan.
2. The question implies that hunting has had a significant impact on the
population of the species in question, but there is no evidence to suggest that
Siberian cranes have declined significantly due to hunting in either Afghanistan
or Asia. In fact, Siberian cranes are considered Near Threatened by the
International Union for Conservation of Nature (IUCN), primarily due to habitat
loss and degradation, not hunting.
3. The humor likely lies in the absurdity of the question and answer pair. It is
unlikely that anyone would mistake Afghanistan as a habitat for Siberian cranes
or believe that hunting has had a significant impact on their population in that
country.
Therefore, based on these factors, it is more likely that the question and
answer pair is meant as humor rather than a factual statement.

_____________MODEL GENERATING RESPONSE_____________


llama_print_timings:        load time =    2587.06 ms
llama_print_timings:      sample time =      13.61 ms /    88 runs   (    0.15
ms per token,  6465.83 tokens per second)
llama_print_timings: prompt eval time =    1067.92 ms /    19 tokens (   56.21
ms per token,    17.79 tokens per second)
llama_print_timings:        eval time =    8623.23 ms /    87 runs   (   99.12
ms per token,    10.09 tokens per second)
llama_print_timings:       total time =    9875.43 ms /   106 tokens

_________________RESPONSE GENERATED_________________

Response Number  25
  This is a humorous answer. The question is asking if February can march, which
makes no sense as February is a month and cannot physically move or perform
actions. The answer, "No, but April may," is a play on words, using the similar
sounding words "February" and "April" to create a pun. It is not a factual
answer, but rather a humorous one.

_________________Local LLM execution complete!_________________

Responses from Local LLM normalized to true and false answers:
 [True, True, True, True, True, True, True, True, True, True, True, True, True,
True, True, True, True, True, True, True, True, True, True, True, True]

Local LLM model llama-2-7b identified 12 out of 25 question-answer pairs
correctly. Accuracy = 48.00%
```