In [2]:
```
!pip install --upgrade pip
!pip install -q transformers ipywidgets huggingface_hub
```

```
Requirement already satisfied: pip in /opt/conda/lib/python3.10/site-packages (23.2.1)
Collecting pip
  Obtaining dependency information for pip from https://files.pythonhosted.org/packages/e7/54/0c1c068542cee73d8863336e974fc881e608d0170f3af15d0c0f28644531/pip-24.1.2-py3-none-any.whl.metadata
  Downloading pip-24.1.2-py3-none-any.whl.metadata (3.6 kB)
Downloading pip-24.1.2-py3-none-any.whl (1.8 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.8/1.8 MB 20.8 MB/s eta 0:00:0000:0100:01
Installing collected packages: pip
  Attempting uninstall: pip
    Found existing installation: pip 23.2.1
    Uninstalling pip-23.2.1:
      Successfully uninstalled pip-23.2.1
Successfully installed pip-24.1.2
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager, possibly rendering your system unusable.It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv. Use the --root-user-action option if you know what you are doing and want to suppress this warning.
```

In [3]:
```
from huggingface_hub import notebook_login
notebook_login()
```

```
VBox(children=(HTML(value='<center> <img\nsrc=https://huggingface.co/front/assets/huggingface_logo-noborder.sv…
```

In [4]:
```
from transformers import AutoTokenizer, AutoModelForCausalLM
```

In [5]:
```
tokenizer = AutoTokenizer.from_pretrained("Jacaranda/UlizaLlama3")
model = AutoModelForCausalLM.from_pretrained("Jacaranda/UlizaLlama3")
```

```
tokenizer_config.json:   0%|          | 0.00/50.6k [00:00<?, ?B/s]
tokenizer.json:   0%|          | 0.00/9.09M [00:00<?, ?B/s]
special_tokens_map.json:   0%|          | 0.00/449 [00:00<?, ?B/s]
config.json:   0%|          | 0.00/713 [00:00<?, ?B/s]
model.safetensors.index.json:   0%|          | 0.00/23.9k [00:00<?, ?B/s]
Downloading shards:   0%|          | 0/4 [00:00<?, ?it/s]
model-00001-of-00004.safetensors:   0%|          | 0.00/4.98G [00:00<?, ?B/s]
model-00002-of-00004.safetensors:   0%|          | 0.00/5.00G [00:00<?, ?B/s]
model-00003-of-00004.safetensors:   0%|          | 0.00/4.92G [00:00<?, ?B/s]
model-00004-of-00004.safetensors:   0%|          | 0.00/1.17G [00:00<?, ?B/s]
Loading checkpoint shards:   0%|          | 0/4 [00:00<?, ?it/s]
generation_config.json:   0%|          | 0.00/172 [00:00<?, ?B/s]
```

In [6]:
```
model
```

Out[6]:
```
LlamaForCausalLM(
  (model): LlamaModel(
    (embed_tokens): Embedding(128256, 4096)
    (layers): ModuleList(
      (0-31): 32 x LlamaDecoderLayer(
        (self_attn): LlamaAttention(
          (q_proj): Linear(in_features=4096, out_features=4096, bias=False)
          (k_proj): Linear(in_features=4096, out_features=1024, bias=False)
          (v_proj): Linear(in_features=4096, out_features=1024, bias=False)
          (o_proj): Linear(in_features=4096, out_features=4096, bias=False)
          (rotary_emb): LlamaRotaryEmbedding()
        )
        (mlp): LlamaMLP(
          (gate_proj): Linear(in_features=4096, out_features=14336, bias=False)
          (up_proj): Linear(in_features=4096, out_features=14336, bias=False)
          (down_proj): Linear(in_features=14336, out_features=4096, bias=False)
          (act_fn): SiLU()
        )
        (input_layernorm): LlamaRMSNorm()
        (post_attention_layernorm): LlamaRMSNorm()
      )
    )
    (norm): LlamaRMSNorm()
    (rotary_emb): LlamaRotaryEmbedding()
  )
  (lm_head): Linear(in_features=4096, out_features=128256, bias=False)
)
```

In [7]:
```
tokenizer
```

Out[7]:
```
PreTrainedTokenizerFast(name_or_path='Jacaranda/UlizaLlama3', vocab_size=128000, model_max_length=1000000000000000019884624838
656, is_fast=True, padding_side='right', truncation_side='right', special_tokens={'bos_token': '<|begin_of_text|>', 'eos_toke
n': '<|end_of_text|>', 'pad_token': '<|end_of_text|>'}, clean_up_tokenization_spaces=True),  added_tokens_decoder={
        128000: AddedToken("<|begin_of_text|>", rstrip=False, lstrip=False, single_word=False, normalized=False, special=Tru
e),
        128001: AddedToken("<|end_of_text|>", rstrip=False, lstrip=False, single_word=False, normalized=False, special=True),
        128002: AddedToken("<|reserved_special_token_0|>", rstrip=False, lstrip=False, single_word=False, normalized=False, sp
ecial=True),
        128003: AddedToken("<|reserved_special_token_1|>", rstrip=False, lstrip=False, single_word=False, normalized=False, sp
ecial=True),
        128004: AddedToken("<|reserved_special_token_2|>", rstrip=False, lstrip=False, single_word=False, normalized=False, sp
ecial=True),
        128005: AddedToken("<|reserved_special_token_3|>", rstrip=False, lstrip=False, single_word=False, normalized=False, sp
ecial=True),
        128006: AddedToken("<|start_header_id|>", rstrip=False, lstrip=False, single_word=False, normalized=False, special=Tru
e),
        128007: AddedToken("<|end_header_id|>", rstrip=False, lstrip=False, single_word=False, normalized=False, special=Tru
e),
        128008: AddedToken("<|reserved_special_token_4|>", rstrip=False, lstrip=False, single_word=False, normalized=False, sp
ecial=True),
        128009: AddedToken("<|eot_id|>", rstrip=False, lstrip=False, single_word=False, normalized=False, special=True),
        128010: AddedToken("<|reserved_special_token_5|>", rstrip=False, lstrip=False, single_word=False, normalized=False, sp
ecial=True),
        128011: AddedToken("<|reserved_special_token_6|>", rstrip=False, lstrip=False, single_word=False, normalized=False, sp
ecial=True),
        128012: AddedToken("<|reserved_special_token_7|>", rstrip=False, lstrip=False, single_word=False, normalized=False, sp
ecial=True),
        128013: AddedToken("<|reserved_special_token_8|>", rstrip=False, lstrip=False, single_word=False, normalized=False, sp
ecial=True),
        128014: AddedToken("<|reserved_special_token_9|>", rstrip=False, lstrip=False, single_word=False, normalized=False, sp
ecial=True),
        128015: AddedToken("<|reserved_special_token_10|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128016: AddedToken("<|reserved_special_token_11|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128017: AddedToken("<|reserved_special_token_12|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128018: AddedToken("<|reserved_special_token_13|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128019: AddedToken("<|reserved_special_token_14|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128020: AddedToken("<|reserved_special_token_15|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128021: AddedToken("<|reserved_special_token_16|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128022: AddedToken("<|reserved_special_token_17|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
```

pecial=True),
        128023: AddedToken("<|reserved_special_token_18|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128024: AddedToken("<|reserved_special_token_19|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128025: AddedToken("<|reserved_special_token_20|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128026: AddedToken("<|reserved_special_token_21|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128027: AddedToken("<|reserved_special_token_22|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128028: AddedToken("<|reserved_special_token_23|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128029: AddedToken("<|reserved_special_token_24|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128030: AddedToken("<|reserved_special_token_25|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128031: AddedToken("<|reserved_special_token_26|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128032: AddedToken("<|reserved_special_token_27|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128033: AddedToken("<|reserved_special_token_28|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128034: AddedToken("<|reserved_special_token_29|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128035: AddedToken("<|reserved_special_token_30|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128036: AddedToken("<|reserved_special_token_31|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128037: AddedToken("<|reserved_special_token_32|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128038: AddedToken("<|reserved_special_token_33|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128039: AddedToken("<|reserved_special_token_34|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128040: AddedToken("<|reserved_special_token_35|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128041: AddedToken("<|reserved_special_token_36|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128042: AddedToken("<|reserved_special_token_37|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128043: AddedToken("<|reserved_special_token_38|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128044: AddedToken("<|reserved_special_token_39|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128045: AddedToken("<|reserved_special_token_40|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s

pecial=True),
        128046: AddedToken("<|reserved_special_token_41|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128047: AddedToken("<|reserved_special_token_42|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128048: AddedToken("<|reserved_special_token_43|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128049: AddedToken("<|reserved_special_token_44|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128050: AddedToken("<|reserved_special_token_45|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128051: AddedToken("<|reserved_special_token_46|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128052: AddedToken("<|reserved_special_token_47|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128053: AddedToken("<|reserved_special_token_48|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128054: AddedToken("<|reserved_special_token_49|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128055: AddedToken("<|reserved_special_token_50|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128056: AddedToken("<|reserved_special_token_51|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128057: AddedToken("<|reserved_special_token_52|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128058: AddedToken("<|reserved_special_token_53|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128059: AddedToken("<|reserved_special_token_54|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128060: AddedToken("<|reserved_special_token_55|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128061: AddedToken("<|reserved_special_token_56|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128062: AddedToken("<|reserved_special_token_57|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128063: AddedToken("<|reserved_special_token_58|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128064: AddedToken("<|reserved_special_token_59|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128065: AddedToken("<|reserved_special_token_60|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128066: AddedToken("<|reserved_special_token_61|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128067: AddedToken("<|reserved_special_token_62|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128068: AddedToken("<|reserved_special_token_63|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s

pecial=True),
        128069: AddedToken("<|reserved_special_token_64|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128070: AddedToken("<|reserved_special_token_65|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128071: AddedToken("<|reserved_special_token_66|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128072: AddedToken("<|reserved_special_token_67|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128073: AddedToken("<|reserved_special_token_68|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128074: AddedToken("<|reserved_special_token_69|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128075: AddedToken("<|reserved_special_token_70|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128076: AddedToken("<|reserved_special_token_71|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128077: AddedToken("<|reserved_special_token_72|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128078: AddedToken("<|reserved_special_token_73|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128079: AddedToken("<|reserved_special_token_74|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128080: AddedToken("<|reserved_special_token_75|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128081: AddedToken("<|reserved_special_token_76|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128082: AddedToken("<|reserved_special_token_77|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128083: AddedToken("<|reserved_special_token_78|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128084: AddedToken("<|reserved_special_token_79|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128085: AddedToken("<|reserved_special_token_80|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128086: AddedToken("<|reserved_special_token_81|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128087: AddedToken("<|reserved_special_token_82|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128088: AddedToken("<|reserved_special_token_83|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128089: AddedToken("<|reserved_special_token_84|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128090: AddedToken("<|reserved_special_token_85|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128091: AddedToken("<|reserved_special_token_86|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s

pecial=True),
        128092: AddedToken("<|reserved_special_token_87|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128093: AddedToken("<|reserved_special_token_88|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128094: AddedToken("<|reserved_special_token_89|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128095: AddedToken("<|reserved_special_token_90|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128096: AddedToken("<|reserved_special_token_91|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128097: AddedToken("<|reserved_special_token_92|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128098: AddedToken("<|reserved_special_token_93|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128099: AddedToken("<|reserved_special_token_94|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128100: AddedToken("<|reserved_special_token_95|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128101: AddedToken("<|reserved_special_token_96|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128102: AddedToken("<|reserved_special_token_97|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128103: AddedToken("<|reserved_special_token_98|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128104: AddedToken("<|reserved_special_token_99|>", rstrip=False, lstrip=False, single_word=False, normalized=False, s
pecial=True),
        128105: AddedToken("<|reserved_special_token_100|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128106: AddedToken("<|reserved_special_token_101|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128107: AddedToken("<|reserved_special_token_102|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128108: AddedToken("<|reserved_special_token_103|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128109: AddedToken("<|reserved_special_token_104|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128110: AddedToken("<|reserved_special_token_105|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128111: AddedToken("<|reserved_special_token_106|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128112: AddedToken("<|reserved_special_token_107|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128113: AddedToken("<|reserved_special_token_108|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128114: AddedToken("<|reserved_special_token_109|>", rstrip=False, lstrip=False, single_word=False, normalized=False,

special=True),
        128115: AddedToken("<|reserved_special_token_110|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128116: AddedToken("<|reserved_special_token_111|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128117: AddedToken("<|reserved_special_token_112|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128118: AddedToken("<|reserved_special_token_113|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128119: AddedToken("<|reserved_special_token_114|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128120: AddedToken("<|reserved_special_token_115|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128121: AddedToken("<|reserved_special_token_116|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128122: AddedToken("<|reserved_special_token_117|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128123: AddedToken("<|reserved_special_token_118|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128124: AddedToken("<|reserved_special_token_119|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128125: AddedToken("<|reserved_special_token_120|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128126: AddedToken("<|reserved_special_token_121|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128127: AddedToken("<|reserved_special_token_122|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128128: AddedToken("<|reserved_special_token_123|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128129: AddedToken("<|reserved_special_token_124|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128130: AddedToken("<|reserved_special_token_125|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128131: AddedToken("<|reserved_special_token_126|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128132: AddedToken("<|reserved_special_token_127|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128133: AddedToken("<|reserved_special_token_128|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128134: AddedToken("<|reserved_special_token_129|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128135: AddedToken("<|reserved_special_token_130|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128136: AddedToken("<|reserved_special_token_131|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128137: AddedToken("<|reserved_special_token_132|>", rstrip=False, lstrip=False, single_word=False, normalized=False,

special=True),
        128138: AddedToken("<|reserved_special_token_133|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128139: AddedToken("<|reserved_special_token_134|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128140: AddedToken("<|reserved_special_token_135|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128141: AddedToken("<|reserved_special_token_136|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128142: AddedToken("<|reserved_special_token_137|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128143: AddedToken("<|reserved_special_token_138|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128144: AddedToken("<|reserved_special_token_139|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128145: AddedToken("<|reserved_special_token_140|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128146: AddedToken("<|reserved_special_token_141|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128147: AddedToken("<|reserved_special_token_142|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128148: AddedToken("<|reserved_special_token_143|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128149: AddedToken("<|reserved_special_token_144|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128150: AddedToken("<|reserved_special_token_145|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128151: AddedToken("<|reserved_special_token_146|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128152: AddedToken("<|reserved_special_token_147|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128153: AddedToken("<|reserved_special_token_148|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128154: AddedToken("<|reserved_special_token_149|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128155: AddedToken("<|reserved_special_token_150|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128156: AddedToken("<|reserved_special_token_151|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128157: AddedToken("<|reserved_special_token_152|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128158: AddedToken("<|reserved_special_token_153|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128159: AddedToken("<|reserved_special_token_154|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128160: AddedToken("<|reserved_special_token_155|>", rstrip=False, lstrip=False, single_word=False, normalized=False,

special=True),
        128161: AddedToken("<|reserved_special_token_156|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128162: AddedToken("<|reserved_special_token_157|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128163: AddedToken("<|reserved_special_token_158|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128164: AddedToken("<|reserved_special_token_159|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128165: AddedToken("<|reserved_special_token_160|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128166: AddedToken("<|reserved_special_token_161|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128167: AddedToken("<|reserved_special_token_162|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128168: AddedToken("<|reserved_special_token_163|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128169: AddedToken("<|reserved_special_token_164|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128170: AddedToken("<|reserved_special_token_165|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128171: AddedToken("<|reserved_special_token_166|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128172: AddedToken("<|reserved_special_token_167|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128173: AddedToken("<|reserved_special_token_168|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128174: AddedToken("<|reserved_special_token_169|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128175: AddedToken("<|reserved_special_token_170|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128176: AddedToken("<|reserved_special_token_171|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128177: AddedToken("<|reserved_special_token_172|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128178: AddedToken("<|reserved_special_token_173|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128179: AddedToken("<|reserved_special_token_174|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128180: AddedToken("<|reserved_special_token_175|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128181: AddedToken("<|reserved_special_token_176|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128182: AddedToken("<|reserved_special_token_177|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128183: AddedToken("<|reserved_special_token_178|>", rstrip=False, lstrip=False, single_word=False, normalized=False,

special=True),
        128184: AddedToken("<|reserved_special_token_179|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128185: AddedToken("<|reserved_special_token_180|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128186: AddedToken("<|reserved_special_token_181|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128187: AddedToken("<|reserved_special_token_182|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128188: AddedToken("<|reserved_special_token_183|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128189: AddedToken("<|reserved_special_token_184|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128190: AddedToken("<|reserved_special_token_185|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128191: AddedToken("<|reserved_special_token_186|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128192: AddedToken("<|reserved_special_token_187|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128193: AddedToken("<|reserved_special_token_188|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128194: AddedToken("<|reserved_special_token_189|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128195: AddedToken("<|reserved_special_token_190|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128196: AddedToken("<|reserved_special_token_191|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128197: AddedToken("<|reserved_special_token_192|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128198: AddedToken("<|reserved_special_token_193|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128199: AddedToken("<|reserved_special_token_194|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128200: AddedToken("<|reserved_special_token_195|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128201: AddedToken("<|reserved_special_token_196|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128202: AddedToken("<|reserved_special_token_197|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128203: AddedToken("<|reserved_special_token_198|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128204: AddedToken("<|reserved_special_token_199|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128205: AddedToken("<|reserved_special_token_200|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128206: AddedToken("<|reserved_special_token_201|>", rstrip=False, lstrip=False, single_word=False, normalized=False,

special=True),
        128207: AddedToken("<|reserved_special_token_202|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128208: AddedToken("<|reserved_special_token_203|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128209: AddedToken("<|reserved_special_token_204|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128210: AddedToken("<|reserved_special_token_205|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128211: AddedToken("<|reserved_special_token_206|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128212: AddedToken("<|reserved_special_token_207|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128213: AddedToken("<|reserved_special_token_208|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128214: AddedToken("<|reserved_special_token_209|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128215: AddedToken("<|reserved_special_token_210|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128216: AddedToken("<|reserved_special_token_211|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128217: AddedToken("<|reserved_special_token_212|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128218: AddedToken("<|reserved_special_token_213|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128219: AddedToken("<|reserved_special_token_214|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128220: AddedToken("<|reserved_special_token_215|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128221: AddedToken("<|reserved_special_token_216|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128222: AddedToken("<|reserved_special_token_217|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128223: AddedToken("<|reserved_special_token_218|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128224: AddedToken("<|reserved_special_token_219|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128225: AddedToken("<|reserved_special_token_220|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128226: AddedToken("<|reserved_special_token_221|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128227: AddedToken("<|reserved_special_token_222|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128228: AddedToken("<|reserved_special_token_223|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128229: AddedToken("<|reserved_special_token_224|>", rstrip=False, lstrip=False, single_word=False, normalized=False,

special=True),
        128230: AddedToken("<|reserved_special_token_225|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128231: AddedToken("<|reserved_special_token_226|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128232: AddedToken("<|reserved_special_token_227|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128233: AddedToken("<|reserved_special_token_228|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128234: AddedToken("<|reserved_special_token_229|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128235: AddedToken("<|reserved_special_token_230|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128236: AddedToken("<|reserved_special_token_231|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128237: AddedToken("<|reserved_special_token_232|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128238: AddedToken("<|reserved_special_token_233|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128239: AddedToken("<|reserved_special_token_234|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128240: AddedToken("<|reserved_special_token_235|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128241: AddedToken("<|reserved_special_token_236|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128242: AddedToken("<|reserved_special_token_237|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128243: AddedToken("<|reserved_special_token_238|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128244: AddedToken("<|reserved_special_token_239|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128245: AddedToken("<|reserved_special_token_240|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128246: AddedToken("<|reserved_special_token_241|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128247: AddedToken("<|reserved_special_token_242|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128248: AddedToken("<|reserved_special_token_243|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128249: AddedToken("<|reserved_special_token_244|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128250: AddedToken("<|reserved_special_token_245|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128251: AddedToken("<|reserved_special_token_246|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128252: AddedToken("<|reserved_special_token_247|>", rstrip=False, lstrip=False, single_word=False, normalized=False,

```
special=True),
        128253: AddedToken("<|reserved_special_token_248|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128254: AddedToken("<|reserved_special_token_249|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
        128255: AddedToken("<|reserved_special_token_250|>", rstrip=False, lstrip=False, single_word=False, normalized=False,
special=True),
}
```

In [23]:
```
!nvidia-smi
```

```
huggingface/tokenizers: The current process just got forked, after parallelism has already been used. Disabling parallelism to a
void deadlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(true | false)
Sat Jul 27 21:58:57 2024
+---------------------------------------------------------------------------------------+
| NVIDIA-SMI 545.23.06              Driver Version: 545.23.06    CUDA Version: 12.3     |
|-----------------------------------------+----------------------+----------------------+
| GPU  Name                 Persistence-M | Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |         Memory-Usage | GPU-Util  Compute M. |
|                                         |                      |               MIG M. |
|=========================================+======================+======================|
|   0  Quadro RTX 8000                On  | 00000000:0A:00.0 Off |                    0 |
| N/A   38C    P0              61W / 250W |  30803MiB / 46080MiB |      0%      Default |
|                                         |                      |                  N/A |
+-----------------------------------------+----------------------+----------------------+

+---------------------------------------------------------------------------------------+
| Processes:                                                                            |
|  GPU   GI   CI        PID   Type   Process name                            GPU Memory |
|        ID   ID                                                             Usage      |
|=======================================================================================|
+---------------------------------------------------------------------------------------+
```

In [24]:
```
import torch

device = torch.device("cuda" if torch.cuda.is_available() else "cpu")
device
```

Out[24]:
```
device(type='cuda')
```

In [26]:
```
model.to(device)
```

Out[26]:
```
LlamaForCausalLM(
  (model): LlamaModel(
    (embed_tokens): Embedding(128256, 4096)
    (layers): ModuleList(
      (0-31): 32 x LlamaDecoderLayer(
        (self_attn): LlamaAttention(
          (q_proj): Linear(in_features=4096, out_features=4096, bias=False)
          (k_proj): Linear(in_features=4096, out_features=1024, bias=False)
          (v_proj): Linear(in_features=4096, out_features=1024, bias=False)
          (o_proj): Linear(in_features=4096, out_features=4096, bias=False)
          (rotary_emb): LlamaRotaryEmbedding()
        )
        (mlp): LlamaMLP(
          (gate_proj): Linear(in_features=4096, out_features=14336, bias=False)
          (up_proj): Linear(in_features=4096, out_features=14336, bias=False)
          (down_proj): Linear(in_features=14336, out_features=4096, bias=False)
          (act_fn): SiLU()
        )
        (input_layernorm): LlamaRMSNorm()
        (post_attention_layernorm): LlamaRMSNorm()
      )
    )
    (norm): LlamaRMSNorm()
    (rotary_emb): LlamaRotaryEmbedding()
  )
  (lm_head): Linear(in_features=4096, out_features=128256, bias=False)
)
```

In [27]:
```python
# Use a pipeline as a high-level helper
from transformers import pipeline

pipe = pipeline("question-answering", model = model, tokenizer = tokenizer, device = 0)
```

```
The model 'LlamaForCausalLM' is not supported for question-answering. Supported models are ['AlbertForQuestionAnswering', 'BartForQuestionAnswering', 'BertForQuestionAnswering', 'BigBirdForQuestionAnswering', 'BigBirdPegasusForQuestionAnswering', 'BloomForQuestionAnswering', 'CamembertForQuestionAnswering', 'CanineForQuestionAnswering', 'ConvBertForQuestionAnswering', 'Data2VecTextForQuestionAnswering', 'DebertaForQuestionAnswering', 'DebertaV2ForQuestionAnswering', 'DistilBertForQuestionAnswering', 'ElectraForQuestionAnswering', 'ErnieForQuestionAnswering', 'ErnieMForQuestionAnswering', 'FalconForQuestionAnswering', 'FlaubertForQuestionAnsweringSimple', 'FNetForQuestionAnswering', 'FunnelForQuestionAnswering', 'GPT2ForQuestionAnswering', 'GPTNeoForQuestionAnswering', 'GPTNeoXForQuestionAnswering', 'GPTJForQuestionAnswering', 'IBertForQuestionAnswering', 'LayoutLMv2ForQuestionAnswering', 'LayoutLMv3ForQuestionAnswering', 'LEDForQuestionAnswering', 'LiltForQuestionAnswering', 'LlamaForQuestionAnswering', 'LongformerForQuestionAnswering', 'LukeForQuestionAnswering', 'LxmertForQuestionAnswering', 'MarkupLMForQuestionAnswering', 'MBartForQuestionAnswering', 'MegaForQuestionAnswering', 'MegatronBertForQuestionAnswering', 'MobileBertForQuestionAnswering', 'MPNetForQuestionAnswering', 'MptForQuestionAnswering', 'MraForQuestionAnswering', 'MT5ForQuestionAnswering', 'MvpForQuestionAnswering', 'NezhaForQuestionAnswering', 'NystromformerForQuestionAnswering', 'OPTForQuestionAnswering', 'QDQBertForQuestionAnswering', 'ReformerForQuestionAnswering', 'RemBertForQuestionAnswering', 'RobertaForQuestionAnswering', 'RobertaPreLayerNormForQuestionAnswering', 'RoCBertForQuestionAnswering', 'RoFormerForQuestionAnswering', 'SplinterForQuestionAnswering', 'SqueezeBertForQuestionAnswering', 'T5ForQuestionAnswering', 'UMT5ForQuestionAnswering', 'XLMForQuestionAnsweringSimple', 'XLMRobertaForQuestionAnswering', 'XLMRobertaXLForQuestionAnswering', 'XLNetForQuestionAnsweringSimple', 'XmodForQuestionAnswering', 'YosoForQuestionAnswering'].
```

In [28]:
```
question = "What is the capital of France?"
```

In [29]:
```
# Preprocess the question
inputs = tokenizer(question, return_tensors="pt")
```

In [30]:
```
inputs['attention_mask']
```

Out[30]:
```
tensor([[1, 1, 1, 1, 1, 1, 1, 1]])
```

In [32]:
```
# Move the inputs to the specified device (GPU or CPU)
inputs = {k: v.to(device) for k, v in inputs.items()}
```

In [33]:
```
# Generate an answer
outputs = model.generate(inputs["input_ids"], attention_mask=inputs["attention_mask"])

# royghly 6 minutes on RTX 8000 X1
```

In [34]:
```
# Convert the answer to text
answer = tokenizer.decode(outputs[0].cpu(), skip_special_tokens=True)

print(answer)
```

What is the capital of France?ParisAnswer: The capital of France is Paris.Answer:

The capital of France is Paris.Answer: The ca

```
pital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of Fr
ance is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Par
is.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: 
The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital
 of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France i
s Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Ans
wer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The ca
pital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of Fr
ance is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Par
is.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: 
The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital
 of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France i
s Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Ans
wer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The ca
pital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of Fr
ance is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Par
is.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: 
The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital
 of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France i
s Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Ans
wer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The ca
pital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of Fr
ance is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Par
is.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: 
The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital
 of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France i
s Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Ans
wer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The ca
pital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of Fr
ance is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Par
is.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: 
The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital of France is Paris.Answer: The capital
 of France is Paris.Answer: The capital of France is Paris.Answer: The capital of
```

In [35]:

```python
question = "Ni sawa kulisha mtoto avocado?"
inputs = tokenizer(question, return_tensors="pt")
inputs = {k: v.to(device) for k, v in inputs.items()}
outputs = model.generate(inputs["input_ids"], attention_mask=inputs["attention_mask"])

answer = tokenizer.decode(outputs[0].cpu(), skip_special_tokens=True)
print(answer)

# roughly 6 minutes  minutes on RTX 8000 X1
```

```
Ni sawa kulisha mtoto avocado? Ndio, ni sawa. Avocado ni chanzo kizuri cha vitamini C, folate, na nyuzi za lishe. Pia ina mafuta
yenye afya, ambayo yanaweza kusaidia kuboresha afya ya ngozi na nywele ya mtoto wako. Hakikisha tu kuwa unachanganya mlo wa mtot
o wako na aina mbalimbali za vyakula vyenye lishe, ikiwa ni pamoja na matunda na mboga, nafaka nzima, protini, na maziwa yenye m
afuta kidogo. Kwa habari zaidi kuhusu kulisha mtoto wako, tafadhali jadili na mtaalamu wa afya au mtoa huduma ya afya.iiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
iiiiiiiiiiiiiiiiiiiiii
```

In [37]:

```python
question = "Asiye funzwa na mamake hufunzwa na?"
inputs = tokenizer(question, return_tensors="pt")
inputs = {k: v.to(device) for k, v in inputs.items()}
outputs = model.generate(inputs["input_ids"], attention_mask=inputs["attention_mask"])

answer = tokenizer.decode(outputs[0].cpu(), skip_special_tokens=True)
print(answer)

# roughly 5 minutes  minutes on RTX 8000 X1
```

```
Asiye funzwa na mamake hufunzwa na? Kosa ni nini katika sentensi hii?  The sentence "He who is not taught by his mother is taugh
t by his own mistake." is a proverb which means that one should learn from their mistakes rather than rely on their parents to t
each them. It is not a grammatical sentence but rather a proverb.
```

In [38]:

```python
question = "Mwenda tezi na omo, marejeo ni?"
inputs = tokenizer(question, return_tensors="pt")
inputs = {k: v.to(device) for k, v in inputs.items()}
outputs = model.generate(inputs["input_ids"], attention_mask=inputs["attention_mask"])

answer = tokenizer.decode(outputs[0].cpu(), skip_special_tokens=True)
print(answer)

# roughly 7 minutes  minutes on RTX 8000 X1
```

Mwenda tezi na omo, marejeo ni?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo,
marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nin
i?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tez
i na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, mare
jeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mw
endaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na
omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo n
i nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaj
i tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo,
marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nin
i?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tez
i na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, mare
jeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mw
endaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na
omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo n
i nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaj
i tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo,
marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nin
i?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tez
i na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, mare
jeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mw
endaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na
omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo n
i nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaj
i tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo,
marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nin
i?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tez
i na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, mare
jeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mw
endaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na
omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo n
i nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaj
i tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo,
marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nin
i?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tez
i na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, mare
jeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mw
endaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na
omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo n
i nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaj
i tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo,
marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nin
i?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tez
i na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, mare
jeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mw
endaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na

```
omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo n
i nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaj
i tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo,
marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nin
i?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tez
i na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, mare
jeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mw
endaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na
omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo n
i nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaj
i tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo,
marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nin
i?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tez
i na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, mare
jeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mw
endaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na
omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo n
i nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaj
i tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo,
marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nin
i?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tez
i na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, mare
jeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mw
endaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na
omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo n
i nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaj
i tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo,
marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nin
i?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tez
i na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo ni nini?Mwendaji tezi na omo, marejeo
```

In [ ]: