

```
!pip install vllm==0.3.3
```

```
Requirement already satisfied: huggingface-hub<1.0,>=0.19.3 in /usr/local/lib/python3.10/dist-packages (from transfo
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.10/dist-packages (from transformers>=4.38
Requirement already satisfied: tokenizers<0.19,>=0.14 in /usr/local/lib/python3.10/dist-packages (from transformers>
Requirement already satisfied: safetensors>=0.4.1 in /usr/local/lib/python3.10/dist-packages (from transformers>=4.3
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.10/dist-packages (from transformers>=4.38.0->vll
Collecting starlette<0.38.0,>=0.37.2 (from fastapi->vllm==0.3.3)
  Downloading starlette-0.37.2-py3-none-any.whl (71 kB)
    71.9/71.9 kB 11.8 MB/s eta 0:00:00
Collecting h11>=0.8 (from uvicorn[standard]->vllm==0.3.3)
  Downloading h11-0.14.0-py3-none-any.whl (58 kB)
    58.3/58.3 kB 10.0 MB/s eta 0:00:00
Collecting httptools>=0.5.0 (from uvicorn[standard]->vllm==0.3.3)
  Downloading httptools-0.6.1-cp310-cp310-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux2014_x86_64.whl (341.4 kB)
    341.4/341.4 kB 40.8 MB/s eta 0:00:00
Collecting python-dotenv>=0.13 (from uvicorn[standard]->vllm==0.3.3)
  Downloading python-dotenv-1.0.1-py3-none-any.whl (19 kB)
Collecting uvloop!=0.15.0,!0.15.1,>=0.14.0 (from uvicorn[standard]->vllm==0.3.3)
  Downloading uvloop-0.19.0-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (3.4 MB)
    3.4/3.4 MB 97.9 MB/s eta 0:00:00
Collecting watchfiles>=0.13 (from uvicorn[standard]->vllm==0.3.3)
  Downloading watchfiles-0.21.0-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (1.3 MB)
    1.3/1.3 MB 81.1 MB/s eta 0:00:00
Collecting websockets>=10.4 (from uvicorn[standard]->vllm==0.3.3)
  Downloading websockets-12.0-cp310-cp310-manylinux_2_5_x86_64.manylinux1_x86_64.manylinux2014_x86_64.whl (130.2 kB)
    130.2/130.2 kB 20.0 MB/s eta 0:00:00
Requirement already satisfied: anyio<5,>=3.4.0 in /usr/local/lib/python3.10/dist-packages (from starlette<0.38.0,>=0
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.10/dist-packages (from jinja2->torch==2.1.2
Requirement already satisfied: attrs>=22.2.0 in /usr/local/lib/python3.10/dist-packages (from jsonschema->outlines>
Requirement already satisfied: jsonschema-specifications>=2023.03.6 in /usr/local/lib/python3.10/dist-packages (from
Requirement already satisfied: rpds-py>=0.7.1 in /usr/local/lib/python3.10/dist-packages (from jsonschema->outlines>
Requirement already satisfied: llvmlite<0.42,>=0.41.0dev0 in /usr/local/lib/python3.10/dist-packages (from numba->ou
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from requests->o
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests->outlines>=0.0
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10/dist-packages (from requests->outline
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests->outline
Requirement already satisfied: mpmath>=0.19 in /usr/local/lib/python3.10/dist-packages (from sympy->torch==2.1.2->vl
Requirement already satisfied: sniffio>=1.1 in /usr/local/lib/python3.10/dist-packages (from anyio<5,>=3.4.0->starle
Requirement already satisfied: exceptiongroup in /usr/local/lib/python3.10/dist-packages (from anyio<5,>=3.4.0->starle
Installing collected packages: ninja, websockets, uvloop, triton, python-dotenv, pynvml, nvidia-nvtx-cu12, nvidia-nv
Attempting uninstall: triton
  Found existing installation: triton 2.2.0
  Uninstalling triton-2.2.0:
    Successfully uninstalled triton-2.2.0
Attempting uninstall: cupy-cuda12x
  Found existing installation: cupy-cuda12x 12.2.0
  Uninstalling cupy-cuda12x-12.2.0:
    Successfully uninstalled cupy-cuda12x-12.2.0
Attempting uninstall: torch
  Found existing installation: torch 2.2.1+cu121
  Uninstalling torch-2.2.1+cu121:
    Successfully uninstalled torch-2.2.1+cu121
ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This beha
torchaudio 2.2.1+cu121 requires torch==2.2.1, but you have torch 2.1.2 which is incompatible.
torchttext 0.17.1 requires torch==2.2.1, but you have torch 2.1.2 which is incompatible.
torchvision 0.17.1+cu121 requires torch==2.2.1, but you have torch 2.1.2 which is incompatible.
Successfully installed cupy-cuda12x-12.1.0 diskcache-5.6.3 fastapi-0.110.1 h11-0.14.0 httptools-0.6.1 interregular-0.
```

```
from vllm import LLM
from vllm import SamplingParams
import pandas as pd
```

```
from google.colab import drive
```

```
drive.mount("/content/drive")
```

```
Mounted at /content/drive
```

```

mini_dev_filepath = "/content/drive/MyDrive/BioLaySumm2024_main/data/mini_dataset/"
mini_dev_plos_filename = "PLOS_val_mini_milestone3.jsonl"
mini_dev_elife_filename = "eLife_val_mini_milestone3.jsonl"

full_dev_filepath = "/content/drive/MyDrive/BioLaySumm2024_main/data/full_dev_dataset/"
full_dev_plos_filename = "PLOS_val.jsonl"
full_dev_elife_filename = "eLife_val.jsonl"

test_filepath = "/content/drive/MyDrive/BioLaySumm2024_main/data/test_dataset/"
test_plos_filename = "PLOS_test.jsonl"
test_elife_filename = "eLife_test.jsonl"

def read_jsonl(filepath, filename):
    df = pd.read_json(filepath + filename,
                      orient="records",
                      lines=True
                     )
    return df

mini_plos_df = read_jsonl(mini_dev_filepath, mini_dev_plos_filename)
mini_elife_df = read_jsonl(mini_dev_filepath, mini_dev_elife_filename)

# full_plos_df = read_jsonl(full_dev_filepath, full_dev_plos_filename)
# full_elife_df = read_jsonl(full_dev_filepath, full_dev_elife_filename)

test_plos_df = read_jsonl(test_filepath, test_plos_filename)
test_elife_df = read_jsonl(test_filepath, test_elife_filename)

llm = LLM(model='BioMistral/BioMistral-7B-DARE-AWQ-QGS128-W4-GEMM')
output = llm.generate('What is the capital cit of British Columbia. Keep answer short.')

print(output)

```

```

/usr/local/lib/python3.10/dist-packages/huggingface_hub/utils/_token.py:88: U
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings ta
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to access p
warnings.warn(
config.json: 100%                               887/887 [00:00<00:00, 79.1kB/s]
WARNING 04-04 00:47:22 config.py:193] awq quantization is not fully optimized
INFO 04-04 00:47:22 llm_engine.py:87] Initializing an LLM engine with config:
tokenizer_config.json: 100%                      1.47k/1.47k [00:00<00:00, 104kB/s]
tokenizer.model: 100%                            493k/493k [00:00<00:00, 25.2MB/s]
tokenizer.json: 100%                             1.80M/1.80M [00:00<00:00, 7.55MB/s]
special_tokens_map.json: 100%                    414/414 [00:00<00:00, 30.2kB/s]
INFO 04-04 00:47:35 weight_utils.py:163] Using model weights format ['*.safet
model.safetensors: 100%                         4.15G/4.15G [00:16<00:00, 388MB/s]
INFO 04-04 00:47:57 llm_engine.py:357] # GPU blocks: 14037, # CPU blocks: 204
INFO 04-04 00:47:59 model_runner.py:684] Capturing the model for CUDA graphs.
INFO 04-04 00:47:59 model_runner.py:688] CUDA graphs can take additional 1~3
INFO 04-04 00:48:06 model_runner.py:756] Graph capturing finished in 7 secs.
Processed prompts: 100%|██████████| 1/1 [00:00<00:00, 13.70it/s]
[RequestOutput(request_id=0, prompt='What is the capital cit of British Colum

```

```

def lay_summarize(text, llm=llm):
    """
        summarize a text using a LLM,
        with min_length and max_length are number of tokens limits for the output
    """

    prompt = f"[INST] Simplify and summarize in around 300 words: {text} [/INST]"

    sampling_params = SamplingParams(temperature=0.8, top_p=0.05, max_tokens=1024)

    output = llm.generate(prompt, sampling_params)

    return output[0].outputs[0].text

```

Start coding or [generate](#) with AI.

```
#test
```

```
s = 'The evolutionary origins of the hypoxia-sensitive cells that trigger amniote respiratory reflexes – carotid body glomus cells'
lay_summarize(s, llm)
```

```
Processed prompts: 100%|██████████| 1/1 [00:03<00:00, 3.31s/it]
' The evolutionary origins of hypoxia-sensitive cells that trigger amniote respiratory reflexes, such as carotid body glomus cells and pulmonary neuroendocrine cells (PNECs), are unclear. Researchers have proposed that glomus cells, which are neural crest-derived, are similar to hypoxia-sensitive neuroepithelial cells (NECs) of fish gills, whose embryonic origin is unknown. NECs have also been likened to PNECs, which differentiate in situ within lung airway epithelia. Using genetic lineage-tracing and neural crest-deficient mutants in zebrafish, and physical fate-mapping in frog and lamprey, researchers have found that NECs are not neural crest-derived, but endoderm-derived,
```

```
#use first 20000 characters as input
```

```
mini_plos_df["biomistral_summary"] = mini_plos_df["article"].apply(lambda text: lay_summarize(text))
mini_plos_df.head()
```

Processed prompts: 100%	<div></div>	1/1	[00:03<00:00, 3.58s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.36s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.41s/it]
Processed prompts: 100%	<div></div>	1/1	[00:03<00:00, 3.90s/it]
Processed prompts: 100%	<div></div>	1/1	[00:03<00:00, 3.37s/it]
Processed prompts: 100%	<div></div>	1/1	[00:03<00:00, 3.35s/it]
Processed prompts: 100%	<div></div>	1/1	[00:03<00:00, 3.28s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.20s/it]
Processed prompts: 100%	<div></div>	1/1	[00:11<00:00, 11.94s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.08s/it]
Processed prompts: 100%	<div></div>	1/1	[00:04<00:00, 4.40s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.60s/it]
Processed prompts: 100%	<div></div>	1/1	[00:11<00:00, 11.51s/it]
Processed prompts: 100%	<div></div>	1/1	[00:11<00:00, 11.78s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.25s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.80s/it]
Processed prompts: 100%	<div></div>	1/1	[00:03<00:00, 3.02s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.68s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.13s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.87s/it]
Processed prompts: 100%	<div></div>	1/1	[00:03<00:00, 3.12s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.63s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.79s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.37s/it]
Processed prompts: 100%	<div></div>	1/1	[00:03<00:00, 3.63s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.00s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.28s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.12s/it]
Processed prompts: 100%	<div></div>	1/1	[00:03<00:00, 3.04s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.82s/it]
Processed prompts: 100%	<div></div>	1/1	[00:03<00:00, 3.28s/it]
Processed prompts: 100%	<div></div>	1/1	[00:12<00:00, 12.14s/it]
Processed prompts: 100%	<div></div>	1/1	[00:11<00:00, 11.71s/it]
Processed prompts: 100%	<div></div>	1/1	[00:11<00:00, 11.17s/it]
Processed prompts: 100%	<div></div>	1/1	[00:11<00:00, 11.13s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.29s/it]
Processed prompts: 100%	<div></div>	1/1	[00:11<00:00, 11.86s/it]
Processed prompts: 100%	<div></div>	1/1	[00:11<00:00, 11.94s/it]
Processed prompts: 100%	<div></div>	1/1	[00:03<00:00, 3.30s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.57s/it]
Processed prompts: 100%	<div></div>	1/1	[00:04<00:00, 4.52s/it]
Processed prompts: 100%	<div></div>	1/1	[00:05<00:00, 5.18s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.40s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.66s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.81s/it]
Processed prompts: 100%	<div></div>	1/1	[00:11<00:00, 11.45s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.92s/it]
Processed prompts: 100%	<div></div>	1/1	[00:03<00:00, 3.96s/it]
Processed prompts: 100%	<div></div>	1/1	[00:03<00:00, 3.72s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.79s/it]
Processed prompts: 100%	<div></div>	1/1	[00:01<00:00, 1.88s/it]
Processed prompts: 100%	<div></div>	1/1	[00:03<00:00, 3.77s/it]
Processed prompts: 100%	<div></div>	1/1	[00:11<00:00, 11.63s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.53s/it]
Processed prompts: 100%	<div></div>	1/1	[00:03<00:00, 3.71s/it]
Processed prompts: 100%	<div></div>	1/1	[00:11<00:00, 11.74s/it]

Next steps:

[Generate code with mini_plos_df](#)[View recommended plots](#)

```
mini_elif_df["biomistral_summary"] = mini_elif_df["article"].apply(lambda text: lay_summarize(text))
mini_elif_df.head()
```

Processed prompts: 100%1/1 [00:11<00:00, 11.93s/it]

Processed prompts: 100%1/1 [00:03<00:00, 3.83s/it]

Processed prompts: 100%1/1 [00:12<00:00, 12.93s/it]

Processed prompts: 100%1/1 [00:11<00:00, 11.97s/it]

Processed prompts: 100%1/1 [00:03<00:00, 3.68s/it]

Processed prompts: 100%1/1 [00:04<00:00, 4.26s/it]

Processed prompts: 100%1/1 [00:13<00:00, 13.34s/it]

Processed prompts: 100%1/1 [00:12<00:00, 12.44s/it]

Processed prompts: 100%1/1 [00:12<00:00, 12.23s/it]

Processed prompts: 100%1/1 [00:04<00:00, 4.24s/it]

Processed prompts: 100%1/1 [00:04<00:00, 4.11s/it]

Processed prompts: 100%1/1 [00:11<00:00, 11.72s/it]

Processed prompts: 100%1/1 [00:11<00:00, 11.99s/it]

Processed prompts: 100%1/1 [00:03<00:00, 3.98s/it]

Processed prompts: 100%1/1 [00:03<00:00, 3.15s/it]

Processed prompts: 100%1/1 [00:02<00:00, 2.47s/it]

Processed prompts: 100%1/1 [00:03<00:00, 3.54s/it]

Processed prompts: 100%1/1 [00:12<00:00, 12.23s/it]

Processed prompts: 100%1/1 [00:03<00:00, 3.90s/it]

Processed prompts: 100%1/1 [00:02<00:00, 2.25s/it]

Processed prompts: 100%1/1 [00:11<00:00, 11.56s/it]

Processed prompts: 100%1/1 [00:12<00:00, 12.19s/it]

Processed prompts: 100%1/1 [00:12<00:00, 12.44s/it]

Processed prompts: 100%1/1 [00:03<00:00, 3.67s/it]

	lay_summary	article	headings	keywords	id	biomistral_summary
0	It can take several months , or even years , f...	Mature neural networks synchronize and integra...	[Abstract, Introduction, Results, Discussion, ...]	[neuroscience]	elife-69011-v2	The authors present a novel approach to study...
1	Many of our decisions are made on the basis of...	Many decisions are thought to arise via the ac...	[Abstract, Introduction, Results, Discussion, ...]	[neuroscience]	elife-17688-v1	The authors present a new model of decision m...
2	Oculo-Cerebro-Renal syndrome of Lowe (Lowe sy...	Mutations in the inositol 5-phosphatase OCRL c...	[Abstract, Introduction, Results, Discussion, ...]	[cell biology]	elife-02975-v2	To determine the role of OCRL in clathrin-med...
3	When an embryo develops , its cells must work ...	Gradients of signaling proteins are essential ...	[Abstract, Introduction, Results, Discussion, ...]	[developmental biology]	elife-38137-v3	The authors have generated a new tool to stud...
4	Our genomes contain a	Similarity between two individuals in the	[Abstract, Introduction, Results, ...]	[evolutionary biology,	elife-15266-	The study by Busby et al.

Next steps:

Generate code with mini_elife_df

View recommended plots

```
test_plos_df["biomistral_summary"] = test_plos_df["article"].apply(lambda text: lay_summarize(text))
test_plos_df.head()
```

Processed prompts: 100%	<div></div>	1/1	[00:11<00:00, 11.28s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.84s/it]
Processed prompts: 100%	<div></div>	1/1	[00:01<00:00, 1.95s/it]
Processed prompts: 100%	<div></div>	1/1	[00:03<00:00, 3.08s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.92s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.05s/it]
Processed prompts: 100%	<div></div>	1/1	[00:03<00:00, 3.01s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.22s/it]
Processed prompts: 100%	<div></div>	1/1	[00:03<00:00, 3.25s/it]
Processed prompts: 100%	<div></div>	1/1	[00:11<00:00, 11.21s/it]
Processed prompts: 100%	<div></div>	1/1	[00:03<00:00, 3.24s/it]
Processed prompts: 100%	<div></div>	1/1	[00:03<00:00, 3.37s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.66s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.99s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.25s/it]
Processed prompts: 100%	<div></div>	1/1	[00:03<00:00, 3.27s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.67s/it]
Processed prompts: 100%	<div></div>	1/1	[00:11<00:00, 11.55s/it]
Processed prompts: 100%	<div></div>	1/1	[00:03<00:00, 3.77s/it]
Processed prompts: 100%	<div></div>	1/1	[00:11<00:00, 11.79s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.22s/it]
Processed prompts: 100%	<div></div>	1/1	[00:04<00:00, 4.94s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.92s/it]
Processed prompts: 100%	<div></div>	1/1	[00:03<00:00, 3.46s/it]
Processed prompts: 100%	<div></div>	1/1	[00:05<00:00, 5.11s/it]
Processed prompts: 100%	<div></div>	1/1	[00:03<00:00, 3.39s/it]
Processed prompts: 100%	<div></div>	1/1	[00:04<00:00, 4.72s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.44s/it]
Processed prompts: 100%	<div></div>	1/1	[00:11<00:00, 11.62s/it]
Processed prompts: 100%	<div></div>	1/1	[00:11<00:00, 11.78s/it]
Processed prompts: 100%	<div></div>	1/1	[00:03<00:00, 3.30s/it]
Processed prompts: 100%	<div></div>	1/1	[00:05<00:00, 5.39s/it]
Processed prompts: 100%	<div></div>	1/1	[00:02<00:00, 2.50s/it]
Processed prompts: 100%	<div></div>	1/1	[00:01<00:00, 1.90s/it]
Processed prompts: 100%	<div></div>	1/1	[00:04<00:00, 4.73s/it]
Processed prompts: 100%	<div></div>	1/1	[00:03<00:00, 3.98s/it]

	article	headings	keywords	id	biomistral_summary
0	Lung-resident (LR) mesenchymal stem and stro...	[Abstract, Introduction, Results, Discussion, ...	[immune system, medical conditions, molecular ...	journal.ppat.1009789	The study aimed to investigate the role of lu...
1	Visceral leishmaniasis (VL) is endemic in So...	[Abstract, Introduction, Methods, Results, Dis...	[neonates, clinical laboratory sciences, trans...	journal.pntd.0007992	The study presents data on 133 pregnant women...
2	A high burden of Salmonella enterica subspecie...	[Abstract, Introduction, Methods, Results, Dis...	[pathogens, medical conditions, taxonomy, bact...	journal.pntd.0010704	This study provides valuable insights into th...
3	Severe Acute Respiratory Syndrome Coronavirus-...	[Abstract, Introduction, Results, Discussion, ...	[pathogens, amniotes, medical conditions, bind...	journal.ppat.1010691	The S2-Fc-murine antibody was generated by fu...
4	Many fungal species utilize hydroxyderivatives...	[Abstract, Introduction, Results and discussio...	[taxonomy, proteins, chemistry, genetics, enzy...	journal.pgen.1009815	The C . parapsilicus genome was sequenced usi...