[2026 LG Aimers 8기]

LG의 거대 언어 모델, EXAONE 경량화 해커톤

LG AI Research

배경

- EXAONE 은 Global Frontier 급 Large-scale 모델과 On-Device를 지원하기 위한 Small-scale 모델이 있음
- 랩탑을 위한 2.4B, 스마트폰을 위한 1.2B 모델이 있으나 더 작고 빠른 모델에 대한 요구사항이 있음
- 단순히 파라미터 수를 더 줄이면 메모리와 속도 요건은 만족하나 정확도가 크게 열화됨
- 모델 크기를 줄이고 빠르게 하면서도 정확도를 유지할 수 있는 경량화 기법을 모색하고자 과제를 제안함

경량화 단계

EXAONE-4.0 분석

경량화 적용

추론 엔진 적용

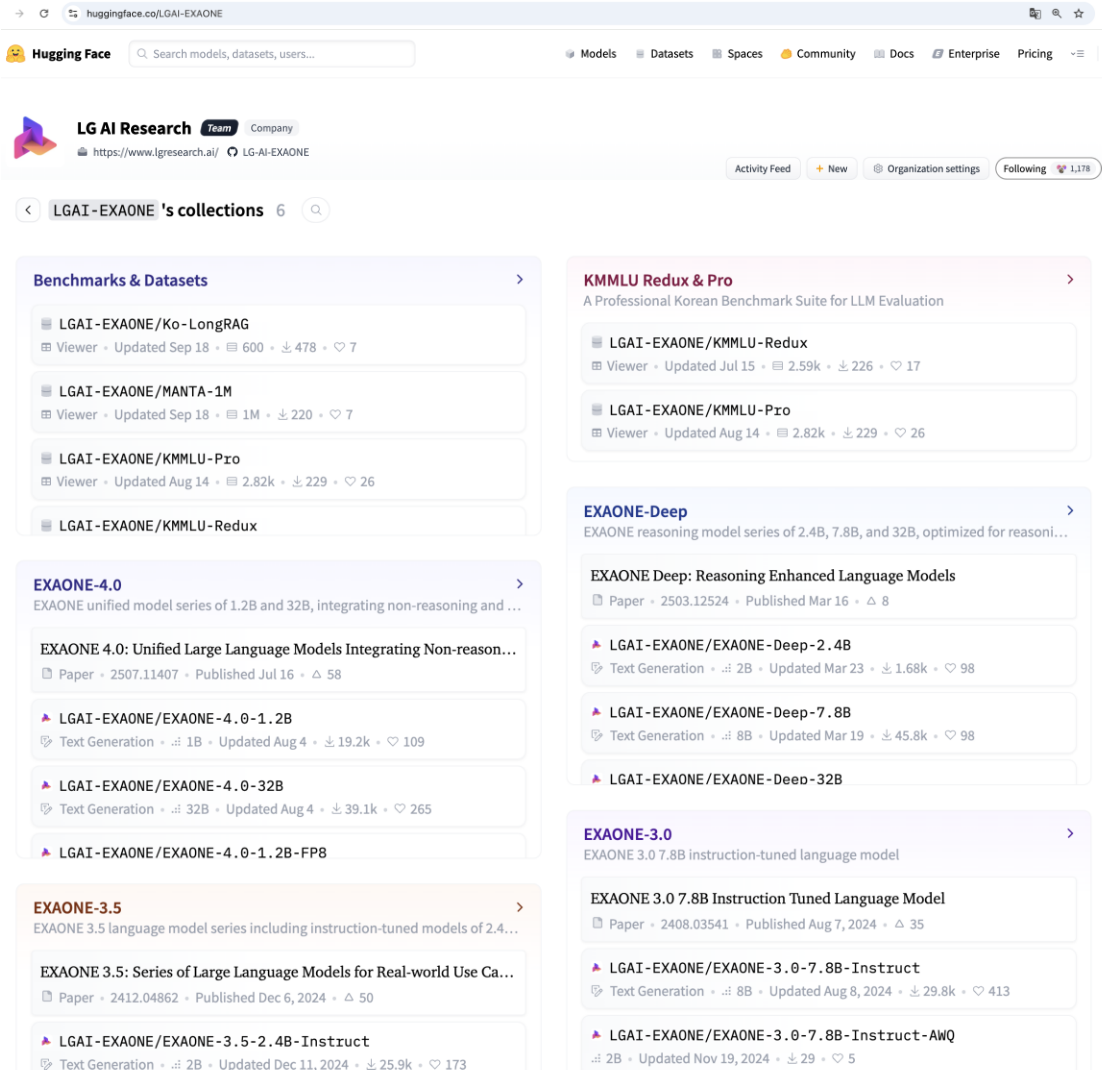
평가

vLLM

기대 효과

- On-Device 환경에서 원활히 구동할 수 있는 EXAONE 모델 지원
- Large-scale EXAONE 모델에도 확대 적용하여 전체 서비스의 운영 비용 감소

EXAONE 4.0 구조



huggingface.co/LGAI-EXAONE

Hugging Face | Search models, datasets, users... | Models | Datasets | Spaces | Community | Docs | Enterprise | Pricing

LG AI Research Team Company
https://www.lgresearch.ai/ LG-AI-EXAONE

Activity Feed | + New | Organization settings | Following 1,178

LGAI-EXAONE 's collections 6

Benchmarks & Datasets

- LGAI-EXAONE/Ko-LongRAG
  Viewer • Updated Sep 18 • 600 • ⤓478 • ♡ 7
- LGAI-EXAONE/MANTA-1M
  Viewer • Updated Sep 18 • 1M • ⤓220 • ♡ 7
- LGAI-EXAONE/KMMLU-Pro
  Viewer • Updated Aug 14 • 2.82k • ⤓229 • ♡ 26
- LGAI-EXAONE/KMMLU-Redux

EXAONE-4.0
EXAONE unified model series of 1.2B and 32B, integrating non-reasoning and ...

- EXAONE 4.0: Unified Large Language Models Integrating Non-reason...
  Paper • 2507.11407 • Published Jul 16 • △ 58
- LGAI-EXAONE/EXAONE-4.0-1.2B
  Text Generation • 1B • Updated Aug 4 • ⤓19.2k • ♡ 109
- LGAI-EXAONE/EXAONE-4.0-32B
  Text Generation • 32B • Updated Aug 4 • ⤓39.1k • ♡ 265
- LGAI-EXAONE/EXAONE-4.0-1.2B-FP8

EXAONE-3.5
EXAONE 3.5 language model series including instruction-tuned models of 2.4...

- EXAONE 3.5: Series of Large Language Models for Real-world Use Ca...
  Paper • 2412.04862 • Published Dec 6, 2024 • △ 50
- LGAI-EXAONE/EXAONE-3.5-2.4B-Instruct
  Text Generation • 2B • Updated Dec 11, 2024 • ⤓25.9k • ♡ 173

KMMLU Redux & Pro
A Professional Korean Benchmark Suite for LLM Evaluation

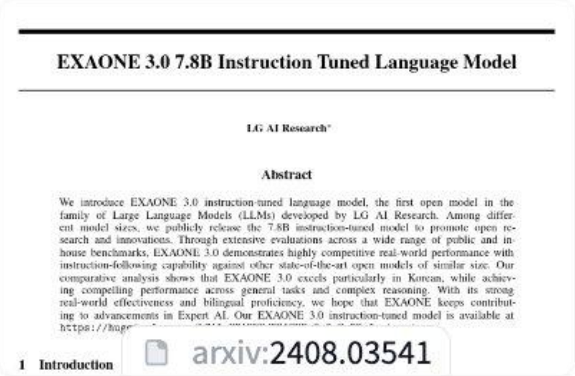
- LGAI-EXAONE/KMMLU-Redux
  Viewer • Updated Jul 15 • 2.59k • ⤓226 • ♡ 17
- LGAI-EXAONE/KMMLU-Pro
  Viewer • Updated Aug 14 • 2.82k • ⤓229 • ♡ 26

EXAONE-Deep
EXAONE reasoning model series of 2.4B, 7.8B, and 32B, optimized for reasoni...

- EXAONE Deep: Reasoning Enhanced Language Models
  Paper • 2503.12524 • Published Mar 16 • △ 8
- LGAI-EXAONE/EXAONE-Deep-2.4B
  Text Generation • 2B • Updated Mar 23 • ⤓1.68k • ♡ 98
- LGAI-EXAONE/EXAONE-Deep-7.8B
  Text Generation • 8B • Updated Mar 19 • ⤓45.8k • ♡ 98
- LGAI-EXAONE/EXAONE-Deep-32B

EXAONE-3.0
EXAONE 3.0 7.8B instruction-tuned language model

- EXAONE 3.0 7.8B Instruction Tuned Language Model
  Paper • 2408.03541 • Published Aug 7, 2024 • △ 35
- LGAI-EXAONE/EXAONE-3.0-7.8B-Instruct
  Text Generation • 8B • Updated Aug 8, 2024 • ⤓29.8k • ♡ 413
- LGAI-EXAONE/EXAONE-3.0-7.8B-Instruct-AWQ
  2B • Updated Nov 19, 2024 • ⤓29 • ♡ 5



EXAONE 3.0 7.8B Instruction Tuned Language Model

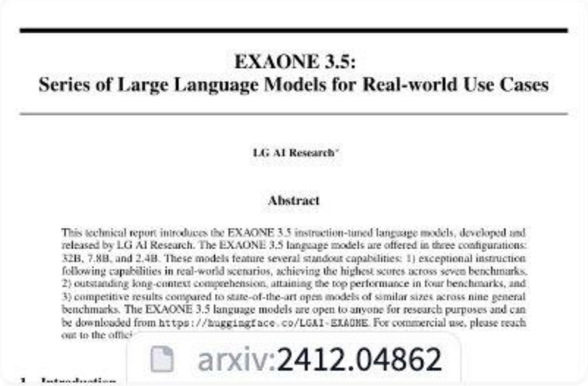
LG AI Research*

Abstract

We introduce EXAONE 3.0 instruction-tuned language model, the first open model in the family of Large Language Models (LLMs) developed by LG AI Research. Among different model sizes, we publicly release the 7.8B instruction-tuned model to promote open research and innovations. Through extensive evaluations across a wide range of public and in-house benchmarks, EXAONE 3.0 demonstrates highly competitive real-world performance with instruction-following capability against other state-of-the-art open models of similar size. Our comparative analysis shows that EXAONE 3.0 excels particularly in Korean, while achieving compelling performance across general tasks and complex reasoning. With its strong real-world effectiveness and bilingual proficiency, we hope that EXAONE keeps contributing to advancements in Expert AI. Our EXAONE 3.0 instruction-tuned model is available at https://hugg

1 Introduction

arxiv:2408.03541



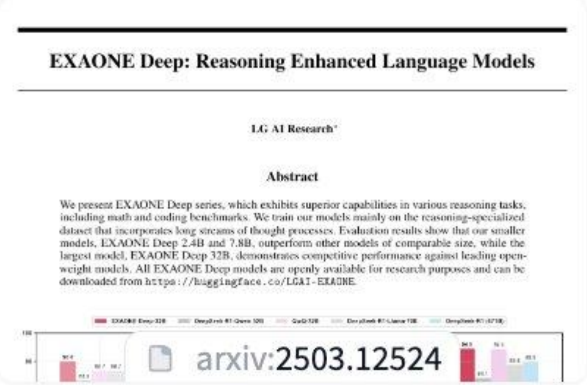
EXAONE 3.5:
Series of Large Language Models for Real-world Use Cases

LG AI Research*

Abstract

This technical report introduces the EXAONE 3.5 instruction-tuned language models, developed and released by LG AI Research. The EXAONE 3.5 language models are offered in three configurations: 32B, 7.8B, and 2.4B. These models feature several standout capabilities: 1) exceptional instruction following capabilities in real-world scenarios, achieving the highest scores across seven benchmarks, 2) outstanding long-context comprehension, attaining the top performance in four benchmarks, and 3) competitive results compared to state-of-the-art open models of similar sizes across nine general benchmarks. The EXAONE 3.5 language models are open to anyone for research purposes and can be downloaded from https://huggingface.co/LGAI-EXAONE. For commercial use, please reach out to the offic

arxiv:2412.04862



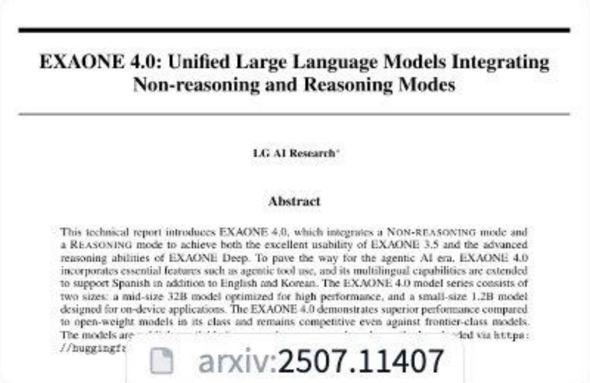
EXAONE Deep: Reasoning Enhanced Language Models

LG AI Research*

Abstract

We present EXAONE Deep series, which exhibits superior capabilities in various reasoning tasks, including math and coding benchmarks. We train our models mainly on the reasoning-specialized dataset that incorporates long streams of thought processes. Evaluation results show that our smaller models, EXAONE Deep 2.4B and 7.8B, outperform other models of comparable size, while the largest model, EXAONE Deep 32B, demonstrates competitive performance against leading open-weight models. All EXAONE Deep models are openly available for research purposes and can be downloaded from https://huggingface.co/LGAI-EXAONE.

arxiv:2503.12524



EXAONE 4.0: Unified Large Language Models Integrating Non-reasoning and Reasoning Modes

LG AI Research*

Abstract

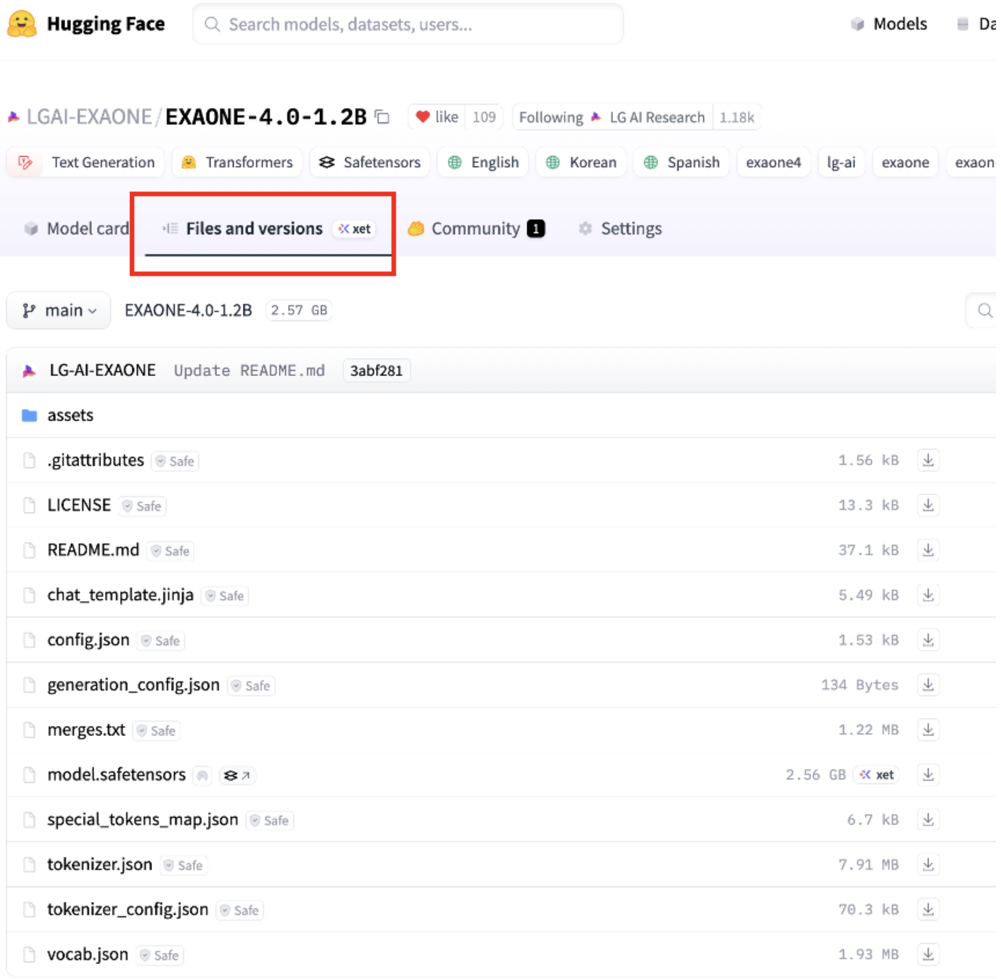
This technical report introduces EXAONE 4.0, which integrates a NON-REASONING mode and a REASONING mode to achieve both the excellent usability of EXAONE 3.5 and the advanced reasoning abilities of EXAONE Deep. To pave the way for the agentic AI era, EXAONE 4.0 incorporates essential features such as agentic tool use, and its multilingual capabilities are extended to support Spanish in addition to English and Korean. The EXAONE 4.0 model series consists of two sizes: a mid-size 32B model optimized for high performance, and a small-size 1.2B model designed for on-device applications. The EXAONE 4.0 demonstrates superior performance compared to open-weight models in its class and remains competitive even against frontier-class models. The models are

ded via https: //huggingf

arxiv:2507.11407

EXAONE 4.0 구조

- 모델 체크포인트는 허깅페이스에서 다운받을 수 있음
- config.json 파일에서 모델의 상세한 구조 정보를 얻을 수 있음

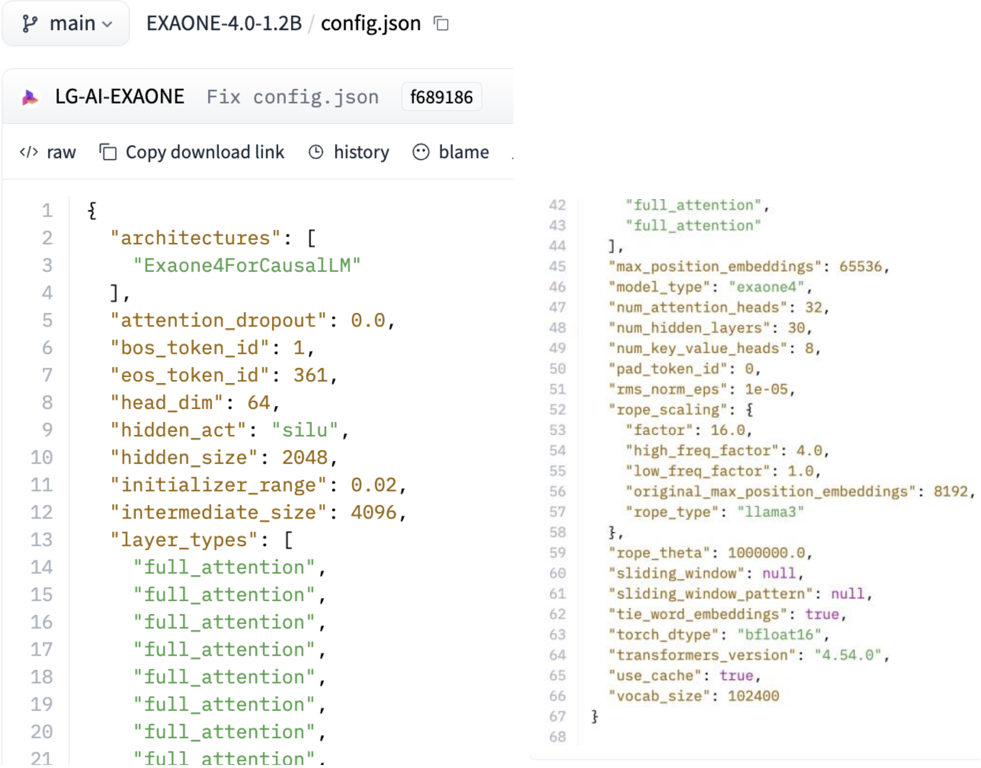


LGAI-EXAONE / **EXAONE-4.0-1.2B** like 109 Following LG AI Research 1.18k

main EXAONE-4.0-1.2B 2.57 GB

LG-AI-EXAONE Update README.md 3abf281

assets		
.gitattributes	Safe	1.56 kB
LICENSE	Safe	13.3 kB
README.md	Safe	37.1 kB
chat_template.jinja	Safe	5.49 kB
config.json	Safe	1.53 kB
generation_config.json	Safe	134 Bytes
merges.txt	Safe	1.22 MB
model.safetensors		2.56 GB xet
special_tokens_map.json	Safe	6.7 kB
tokenizer.json	Safe	7.91 MB
tokenizer_config.json	Safe	70.3 kB
vocab.json	Safe	1.93 MB



main EXAONE-4.0-1.2B / config.json

LG-AI-EXAONE Fix config.json f689186

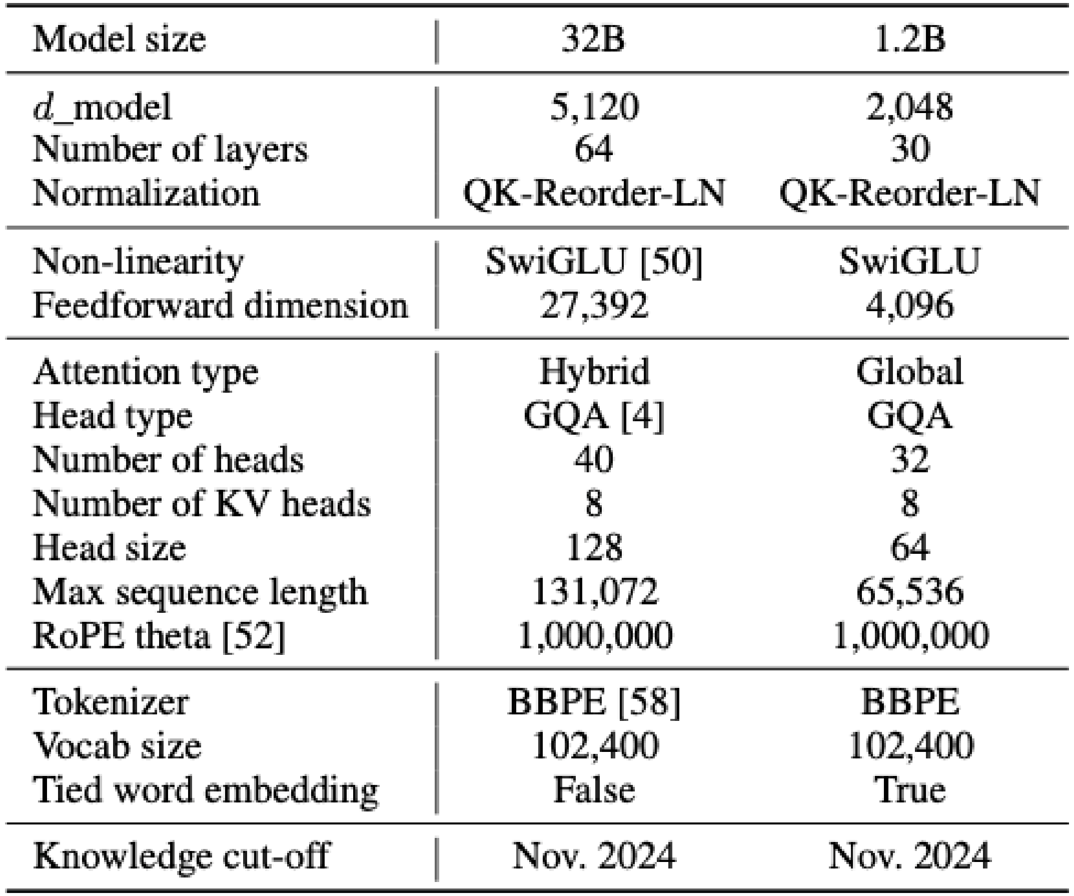
```
{
  "architectures": [
    "Exaone4ForCausalLM"
  ],
  "attention_dropout": 0.0,
  "bos_token_id": 1,
  "eos_token_id": 361,
  "head_dim": 64,
  "hidden_act": "silu",
  "hidden_size": 2048,
  "initializer_range": 0.02,
  "intermediate_size": 4096,
  "layer_types": [
    "full_attention",
    "full_attention",
    "full_attention",
    "full_attention",
    "full_attention",
    "full_attention",
    "full_attention",
    "full_attention",
```

```
    "full_attention",
    "full_attention"
  ],
  "max_position_embeddings": 65536,
  "model_type": "exaone4",
  "num_attention_heads": 32,
  "num_hidden_layers": 30,
  "num_key_value_heads": 8,
  "pad_token_id": 0,
  "rms_norm_eps": 1e-05,
  "rope_scaling": {
    "factor": 16.0,
    "high_freq_factor": 4.0,
    "low_freq_factor": 1.0,
    "original_max_position_embeddings": 8192,
    "rope_type": "llama3"
  },
  "rope_theta": 1000000.0,
  "sliding_window": null,
  "sliding_window_pattern": null,
  "tie_word_embeddings": true,
  "torch_dtype": "bfloat16",
  "transformers_version": "4.54.0",
  "use_cache": true,
  "vocab_size": 102400
}
```

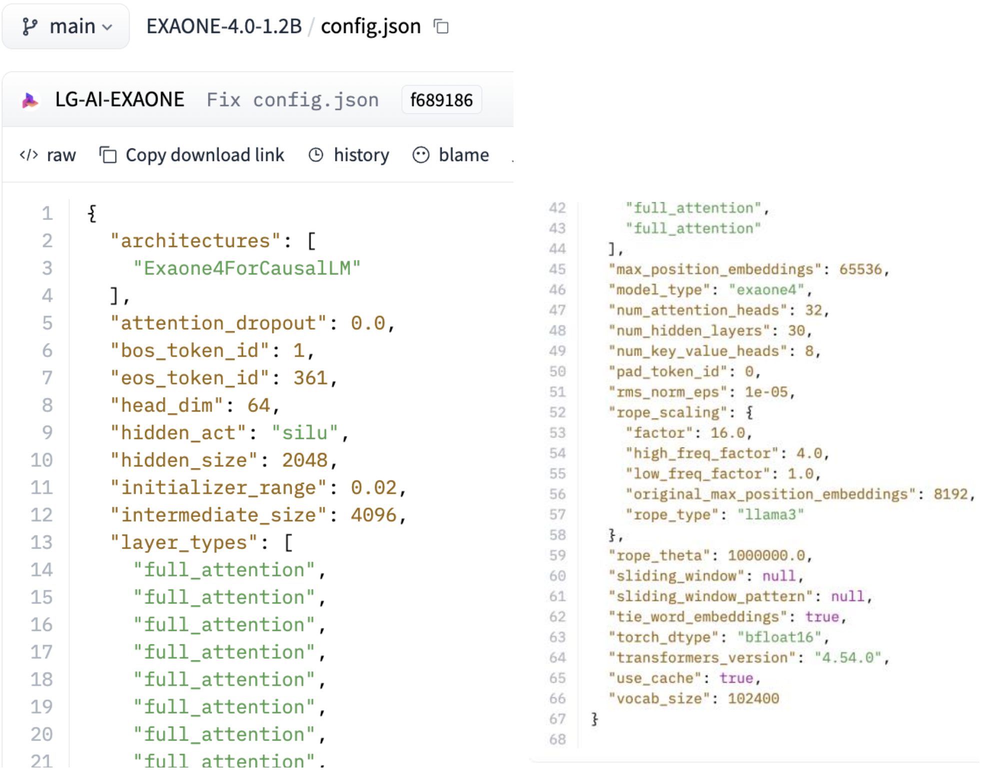
https://huggingface.co/LGAI-EXAONE/EXAONE-4.0-1.2B/tree/main

EXAONE 4.0 구조

- config.json에서 핵심이 되는 부분들은 한눈에 보이게 EXAONE-4.0 Technical Report에 정리되어 있음
- 32B와 1.2B는 파라미터 크기를 제외하고도 Attention Type, Tie word embedding 등이 다름



Model size	32B	1.2B
*d*_model	5,120	2,048
Number of layers	64	30
Normalization	QK-Reorder-LN	QK-Reorder-LN
Non-linearity	SwiGLU [50]	SwiGLU
Feedforward dimension	27,392	4,096
Attention type	Hybrid	Global
Head type	GQA [4]	GQA
Number of heads	40	32
Number of KV heads	8	8
Head size	128	64
Max sequence length	131,072	65,536
RoPE theta [52]	1,000,000	1,000,000
Tokenizer	BBPE [58]	BBPE
Vocab size	102,400	102,400
Tied word embedding	False	True
Knowledge cut-off	Nov. 2024	Nov. 2024



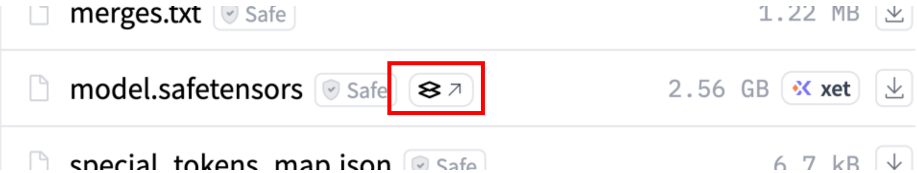
```
{
  "architectures": [
    "Exaone4ForCausalLM"
  ],
  "attention_dropout": 0.0,
  "bos_token_id": 1,
  "eos_token_id": 361,
  "head_dim": 64,
  "hidden_act": "silu",
  "hidden_size": 2048,
  "initializer_range": 0.02,
  "intermediate_size": 4096,
  "layer_types": [
    "full_attention",
    "full_attention",
    "full_attention",
    "full_attention",
    "full_attention",
    "full_attention",
    "full_attention",
    "full_attention",
```

```
    "full_attention",
    "full_attention"
  ],
  "max_position_embeddings": 65536,
  "model_type": "exaone4",
  "num_attention_heads": 32,
  "num_hidden_layers": 30,
  "num_key_value_heads": 8,
  "pad_token_id": 0,
  "rms_norm_eps": 1e-05,
  "rope_scaling": {
    "factor": 16.0,
    "high_freq_factor": 4.0,
    "low_freq_factor": 1.0,
    "original_max_position_embeddings": 8192,
    "rope_type": "llama3"
  },
  "rope_theta": 1000000.0,
  "sliding_window": null,
  "sliding_window_pattern": null,
  "tie_word_embeddings": true,
  "torch_dtype": "bfloat16",
  "transformers_version": "4.54.0",
  "use_cache": true,
  "vocab_size": 102400
}
```

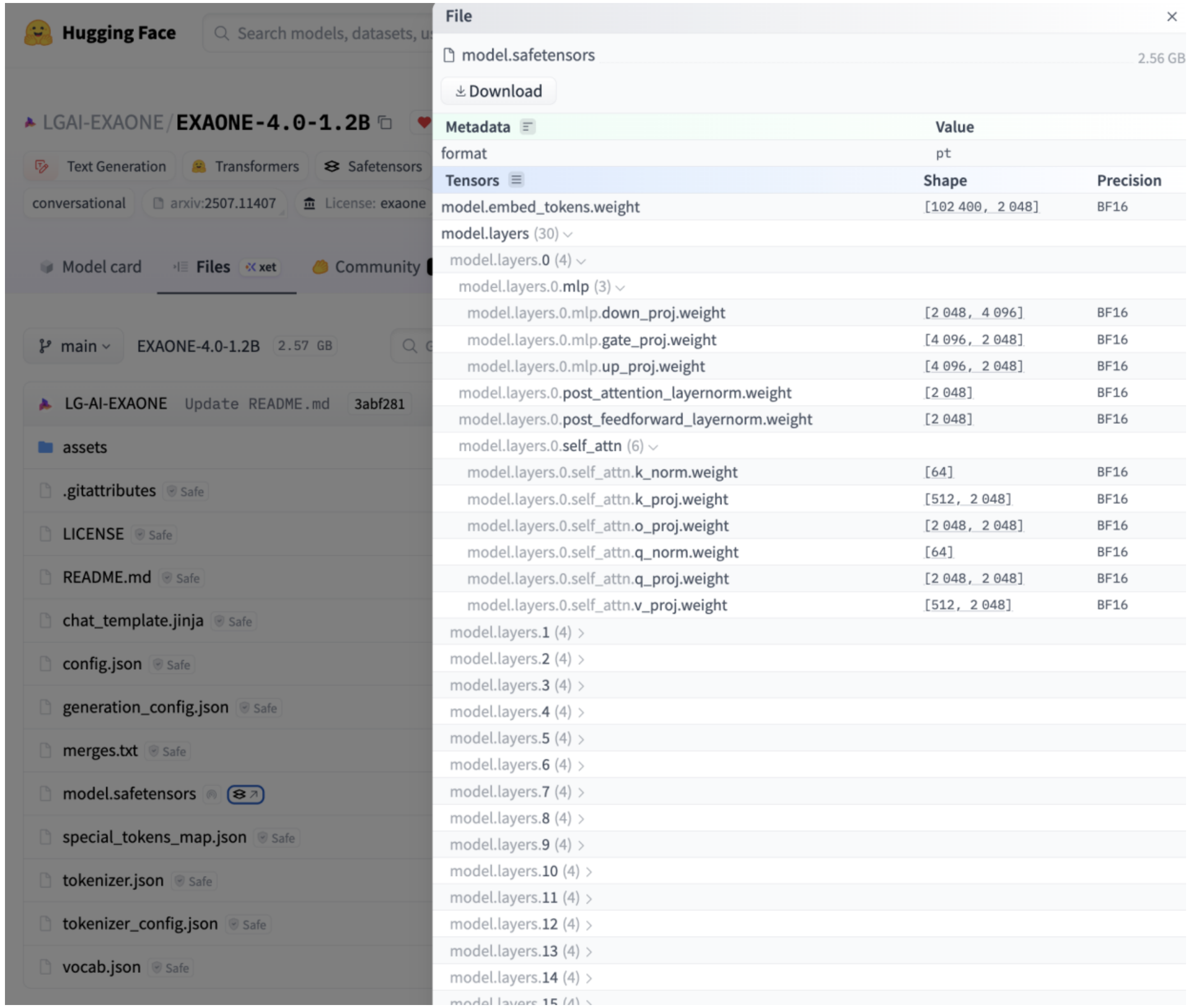
https://arxiv.org/abs/2507.11407

EXAONE 4.0 구조

- 모델의 모듈별 shape, precision 정보를 웹페이지에서도 얻을 수 있음
- 간단한 파라미터 숫자, 용량에 대한 계산에 용이함



merges.txt Safe	1.22 MB	
model.safetensors Safe	2.56 GB	xet
special_tokens_map.json Safe	6.7 kB	



Hugging Face Search models, datasets, u

LGAI-EXAONE / EXAONE-4.0-1.2B

Text Generation · Transformers · Safetensors · conversational · arxiv:2507.11407 · License: exaone

Model card · Files xet · Community

main · EXAONE-4.0-1.2B · 2.57 GB

LG-AI-EXAONE Update README.md 3abf281

- assets
- .gitattributes Safe
- LICENSE Safe
- README.md Safe
- chat_template.jinja Safe
- config.json Safe
- generation_config.json Safe
- merges.txt Safe
- model.safetensors
- special_tokens_map.json Safe
- tokenizer.json Safe
- tokenizer_config.json Safe
- vocab.json Safe

File

model.safetensors 2.56 GB

Download

Metadata	Value
format	pt

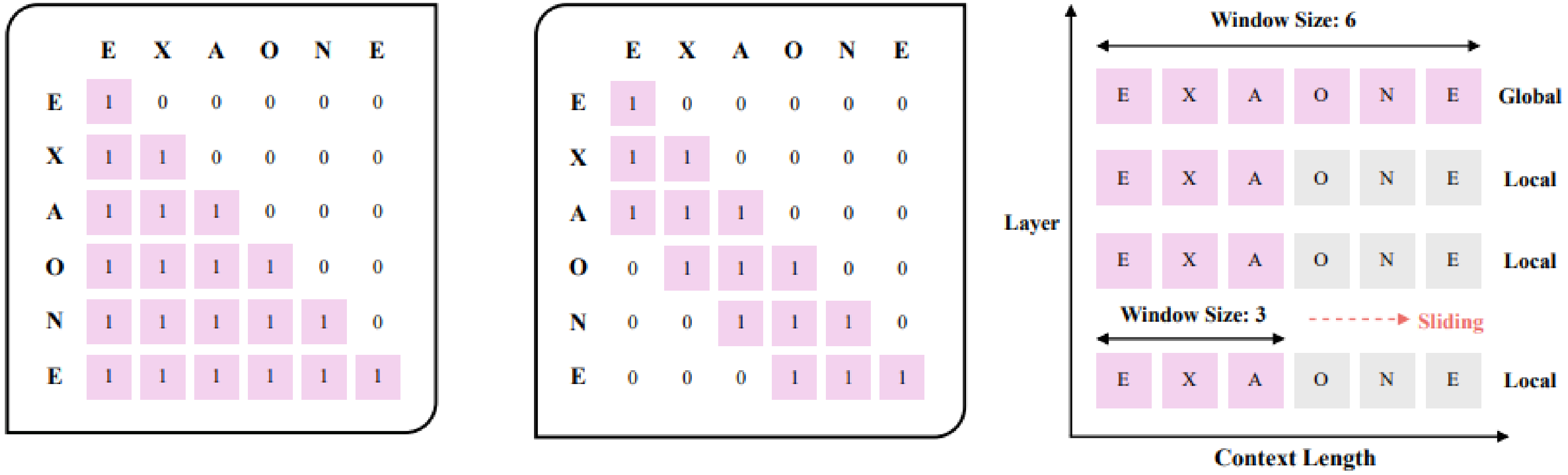
Tensors	Shape	Precision
model.embed_tokens.weight	[102 400, 2 048]	BF16
model.layers (30)		
model.layers.0 (4)		
model.layers.0.mlp (3)		
model.layers.0.mlp.down_proj.weight	[2 048, 4 096]	BF16
model.layers.0.mlp.gate_proj.weight	[4 096, 2 048]	BF16
model.layers.0.mlp.up_proj.weight	[4 096, 2 048]	BF16
model.layers.0.post_attention_layernorm.weight	[2 048]	BF16
model.layers.0.post_feedforward_layernorm.weight	[2 048]	BF16
model.layers.0.self_attn (6)		
model.layers.0.self_attn.k_norm.weight	[64]	BF16
model.layers.0.self_attn.k_proj.weight	[512, 2 048]	BF16
model.layers.0.self_attn.o_proj.weight	[2 048, 2 048]	BF16
model.layers.0.self_attn.q_norm.weight	[64]	BF16
model.layers.0.self_attn.q_proj.weight	[2 048, 2 048]	BF16
model.layers.0.self_attn.v_proj.weight	[512, 2 048]	BF16
model.layers.1 (4)		
model.layers.2 (4)		
model.layers.3 (4)		
model.layers.4 (4)		
model.layers.5 (4)		
model.layers.6 (4)		
model.layers.7 (4)		
model.layers.8 (4)		
model.layers.9 (4)		
model.layers.10 (4)		
model.layers.11 (4)		
model.layers.12 (4)		
model.layers.13 (4)		
model.layers.14 (4)		
model.layers.15 (4)		

https://huggingface.co/LGAI-EXAONE/EXAONE-4.0-1.2B/tree/main?show_file_info=model.safetensors

EXAONE 4.0 구조

EXAONE-4.0의 구조적 특징 두가지 (**Sliding Window Hybrid Attention,** QK-Reorder-LN)

Sliding Window Hybrid Attention (32B)



	E	X	A	O	N	E
E	1	0	0	0	0	0
X	1	1	0	0	0	0
A	1	1	1	0	0	0
O	1	1	1	1	0	0
N	1	1	1	1	1	0
E	1	1	1	1	1	1

Global Attention

	E	X	A	O	N	E
E	1	0	0	0	0	0
X	1	1	0	0	0	0
A	1	1	1	0	0	0
O	0	1	1	1	0	0
N	0	0	1	1	1	0
E	0	0	0	1	1	1

Sliding Window Attention

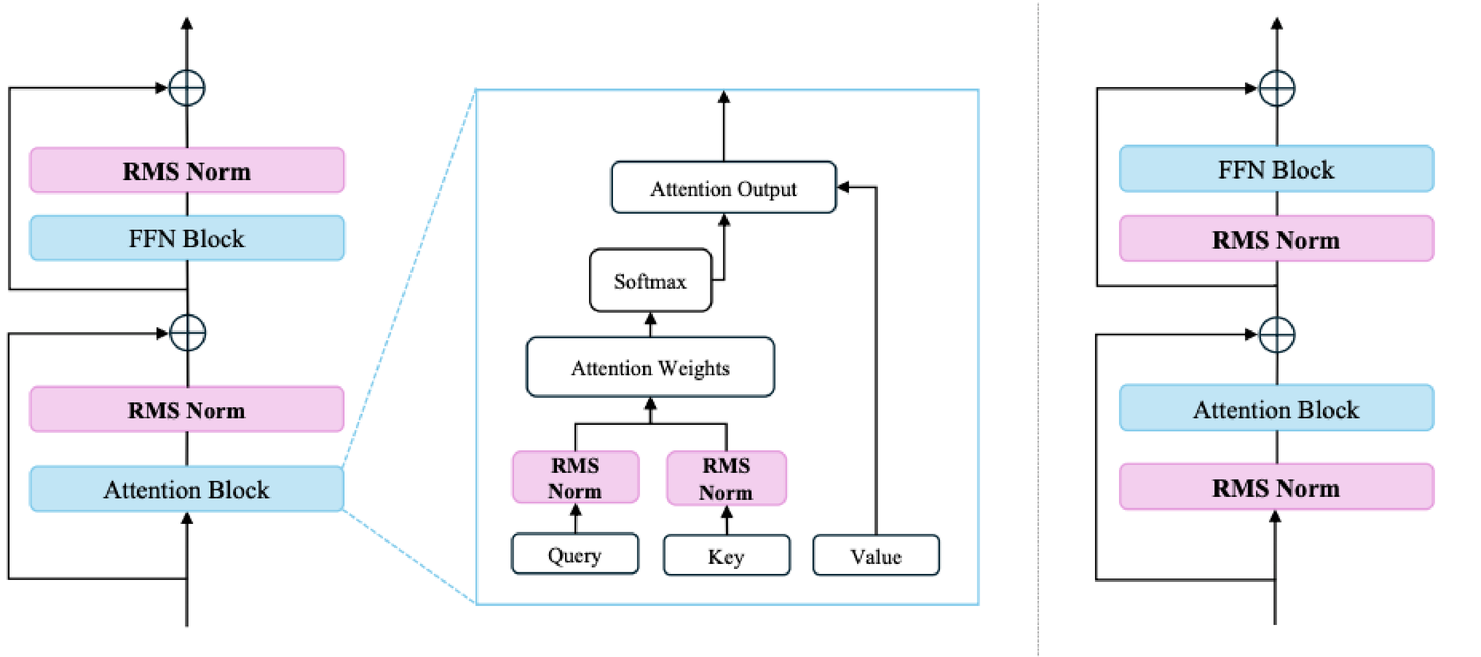
Local Global (LG) Ratio - 3:1

- 3 : 1 비율로 Local (Sliding Window) Attention과 Global Attention을 Hybrid로 적용함
- Local Attention을 적용해 Attention 연산을 줄이고 추론시 KV Cache Memory를 절감함
- Global Attention을 Hybrid로 사용해 열화되는 정확도를 보존함

EXAONE 4.0 구조

EXAONE-4.0의 구조적 특징 두가지 (Sliding Window Hybrid Attention, **QK-Reorder-LN**)

QK-Reorder-LN

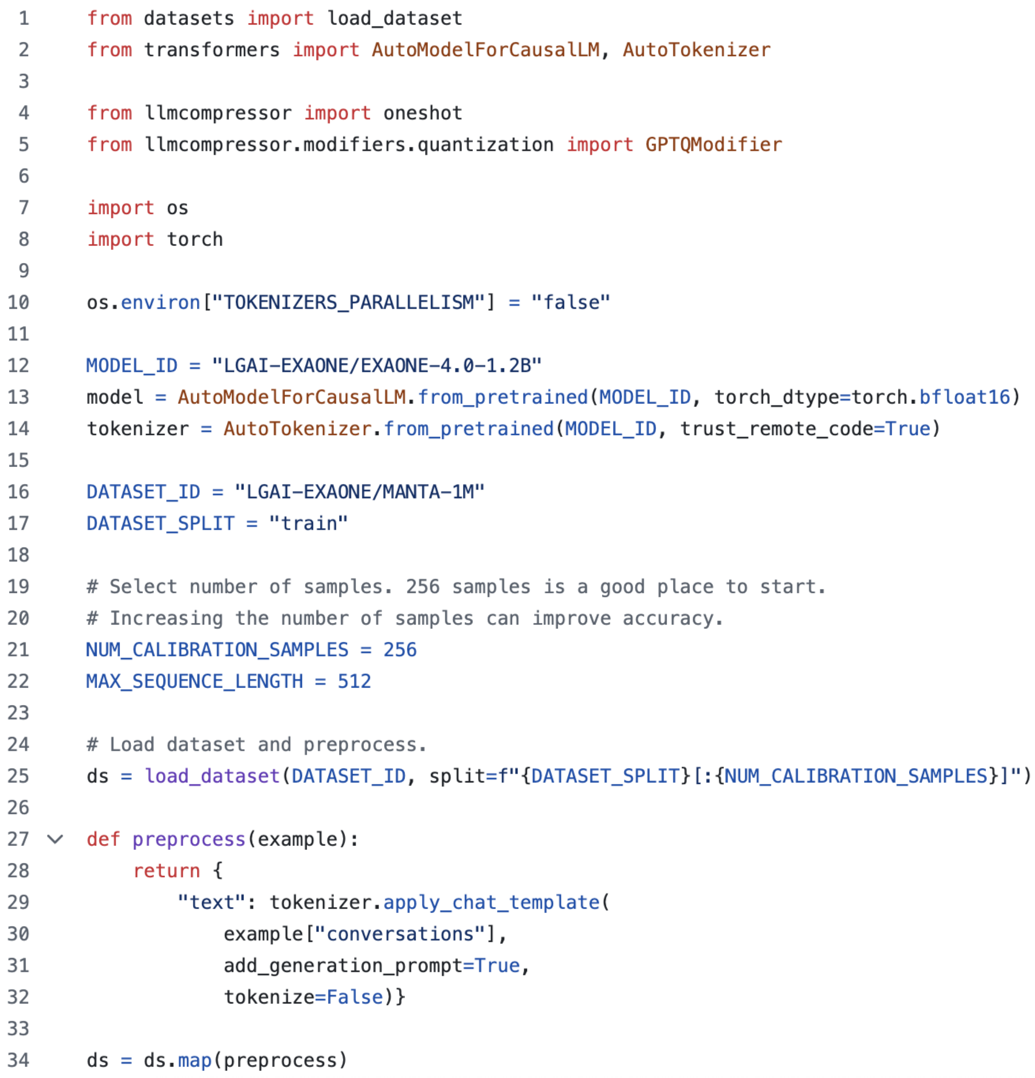


EXAONE 4.0 (QK-Reorder-LN)

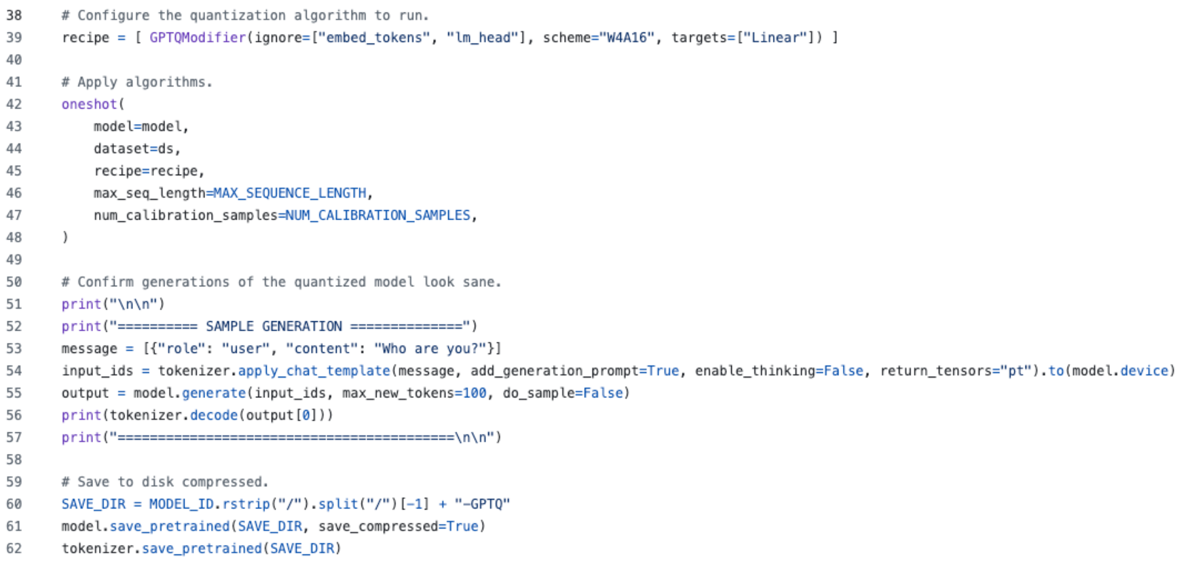
EXAONE 3.5 (Pre-LN)

- LayerNorm의 위치를 변경하고 Query, Key Projection에 LayerNorm을 추가함
- 약간의 연산량 추가로 더 높은 성능을 달성할 수 있음

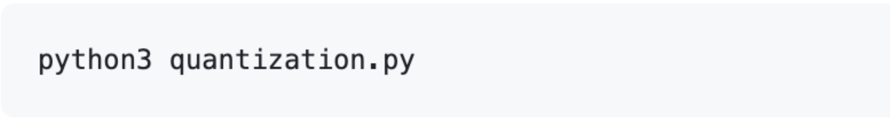
경량화 적용 - LLM Compressor



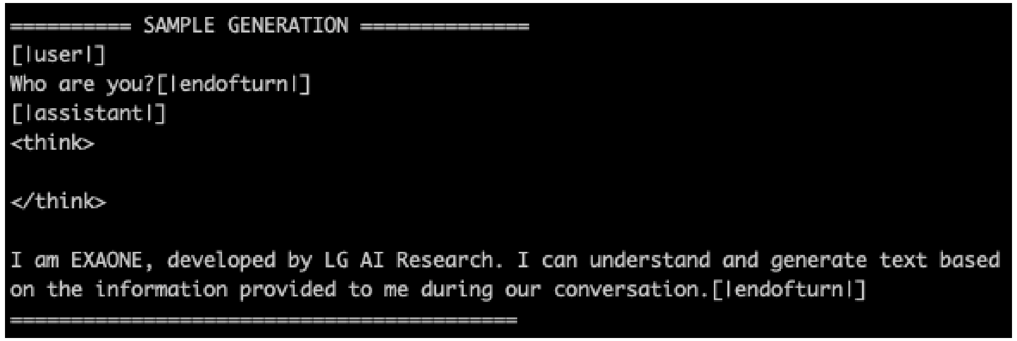
```python
from datasets import load_dataset
from transformers import AutoModelForCausalLM, AutoTokenizer

from llmcompressor import oneshot
from llmcompressor.modifiers.quantization import GPTQModifier

import os
import torch

os.environ["TOKENIZERS_PARALLELISM"] = "false"

MODEL_ID = "LGAI-EXAONE/EXAONE-4.0-1.2B"
model = AutoModelForCausalLM.from_pretrained(MODEL_ID, torch_dtype=torch.bfloat16)
tokenizer = AutoTokenizer.from_pretrained(MODEL_ID, trust_remote_code=True)

DATASET_ID = "LGAI-EXAONE/MANTA-1M"
DATASET_SPLIT = "train"

# Select number of samples. 256 samples is a good place to start.
# Increasing the number of samples can improve accuracy.
NUM_CALIBRATION_SAMPLES = 256
MAX_SEQUENCE_LENGTH = 512

# Load dataset and preprocess.
ds = load_dataset(DATASET_ID, split=f"{DATASET_SPLIT}[:{NUM_CALIBRATION_SAMPLES}]")

def preprocess(example):
    return {
        "text": tokenizer.apply_chat_template(
            example["conversations"],
            add_generation_prompt=True,
            tokenize=False)}

ds = ds.map(preprocess)
```



```python
# Configure the quantization algorithm to run.
recipe = [ GPTQModifier(ignore=["embed_tokens", "lm_head"], scheme="W4A16", targets=["Linear"]) ]

# Apply algorithms.
oneshot(
    model=model,
    dataset=ds,
    recipe=recipe,
    max_seq_length=MAX_SEQUENCE_LENGTH,
    num_calibration_samples=NUM_CALIBRATION_SAMPLES,
)

# Confirm generations of the quantized model look sane.
print("\n\n")
print("========== SAMPLE GENERATION ==============")
message = [{"role": "user", "content": "Who are you?"}]
input_ids = tokenizer.apply_chat_template(message, add_generation_prompt=True, enable_thinking=False, return_tensors="pt").to(model.device)
output = model.generate(input_ids, max_new_tokens=100, do_sample=False)
print(tokenizer.decode(output[0]))
print("==========================================\n\n")

# Save to disk compressed.
SAVE_DIR = MODEL_ID.rstrip("/").split("/")[-1] + "-GPTQ"
model.save_pretrained(SAVE_DIR, save_compressed=True)
tokenizer.save_pretrained(SAVE_DIR)
```

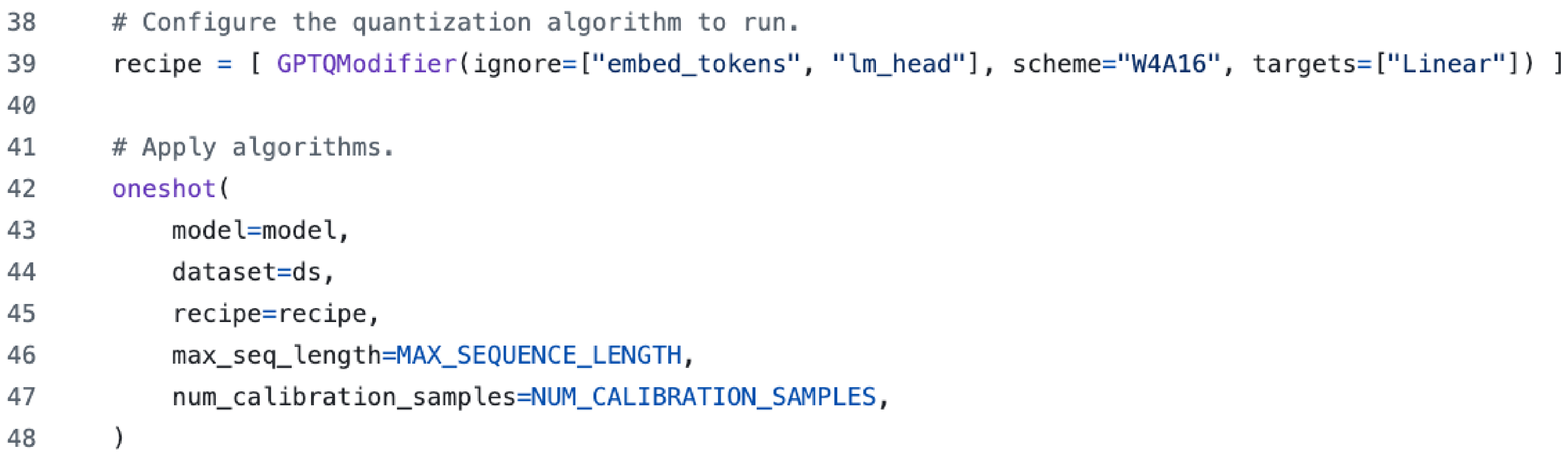


```
python3 quantization.py
```



```
========== SAMPLE GENERATION ==============
[|user|]
Who are you?[|endofturn|]
[|assistant|]
<think>

</think>

I am EXAONE, developed by LG AI Research. I can understand and generate text based
on the information provided to me during our conversation.[|endofturn|]
==========================================
```

경량화 적용 - LLM Compressor



```python
# Configure the quantization algorithm to run.
recipe = [ GPTQModifier(ignore=["embed_tokens", "lm_head"], scheme="W4A16", targets=["Linear"]) ]

# Apply algorithms.
oneshot(
    model=model,
    dataset=ds,
    recipe=recipe,
    max_seq_length=MAX_SEQUENCE_LENGTH,
    num_calibration_samples=NUM_CALIBRATION_SAMPLES,
)
```

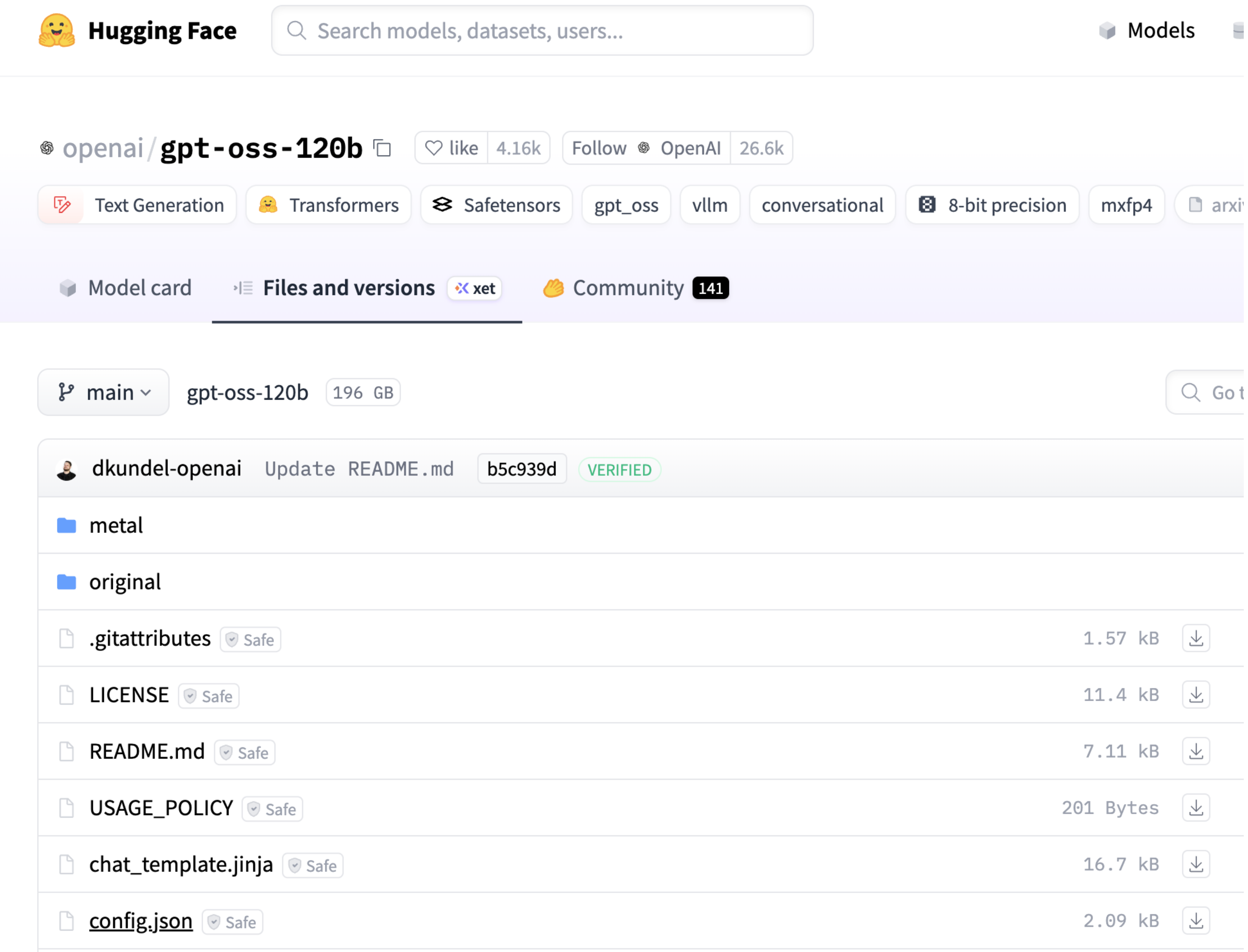
- ignore : 양자화를 제외할 모듈을 지정하는 인자
- scheme : Weight와 Activtation을 어떤 precision으로 사용할지 지정하는 인자
- targets : ignore와 반대로 양자화를 할 모듈을 지정하는 인자

https://github.com/vllm-project/llm-compressor/tree/main/examples

https://github.com/vllm-project/llm-compressor/blob/main/docs/guides/compression_schemes.md

https://github.com/ModelCloud/GPTQModel/tree/main?tab=readme-ov-file#quantization-support

경량화 적용 – 최신 모델 경향



Hugging Face Search models, datasets, users... Models

openai/gpt-oss-120b like 4.16k Follow OpenAI 26.6k

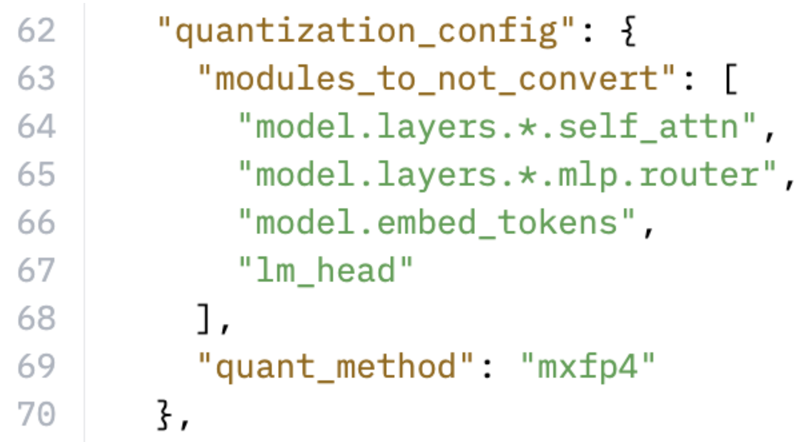
Text Generation · Transformers · Safetensors · gpt_oss · vllm · conversational · 8-bit precision · mxfp4 · arxi

Model card · Files and versions xet · Community 141

main gpt-oss-120b 196 GB Go t

dkundel-openai Update README.md b5c939d VERIFIED

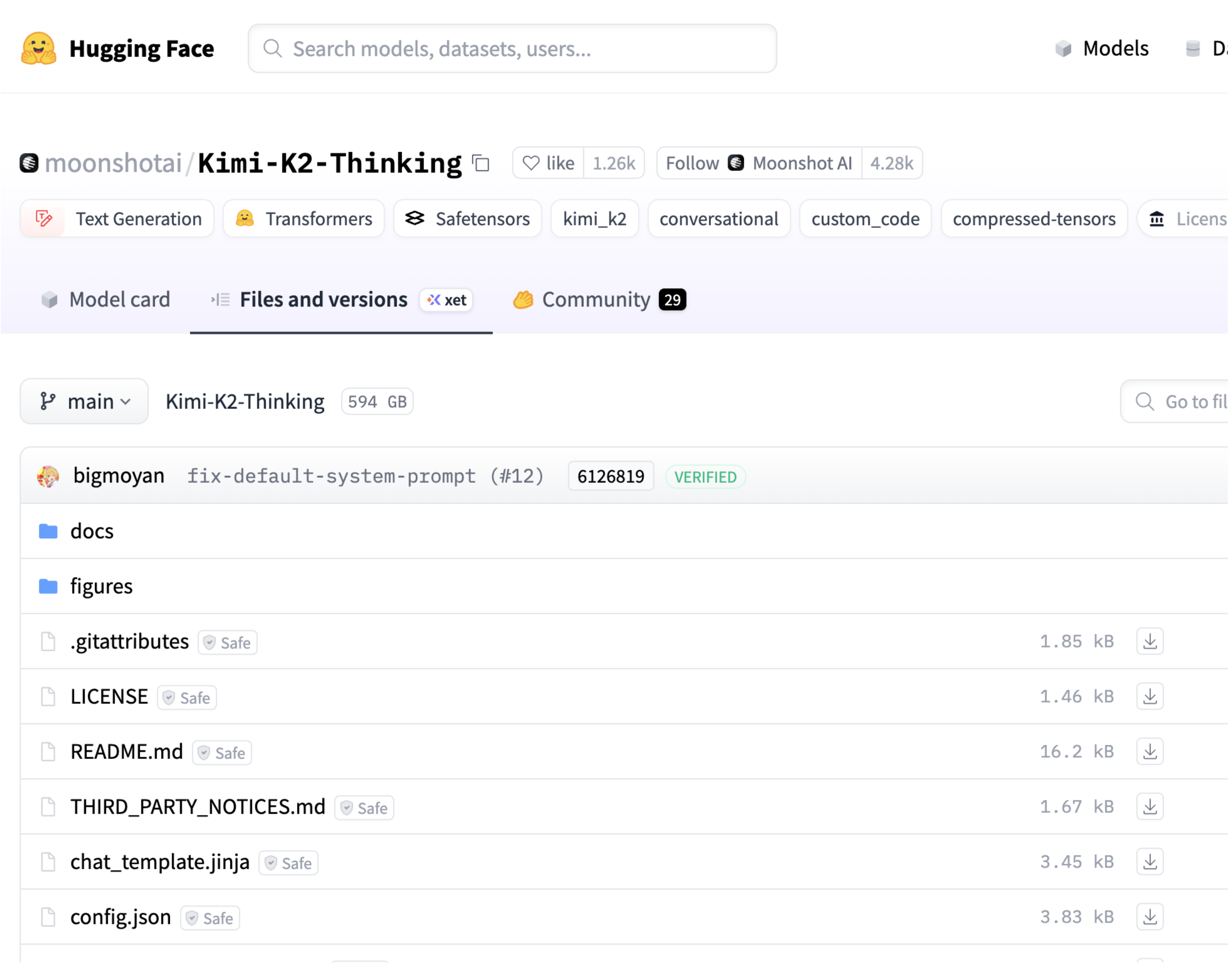
metal	
original	
.gitattributes Safe	1.57 kB
LICENSE Safe	11.4 kB
README.md Safe	7.11 kB
USAGE_POLICY Safe	201 Bytes
chat_template.jinja Safe	16.7 kB
config.json Safe	2.09 kB



```
    "quantization_config": {
      "modules_to_not_convert": [
        "model.layers.*.self_attn",
        "model.layers.*.mlp.router",
        "model.embed_tokens",
        "lm_head"
      ],
      "quant_method": "mxfp4"
    },
```

https://huggingface.co/openai/gpt-oss-120b/blob/main/config.json#L62-L70

경량화 적용 – 최신 모델 경향

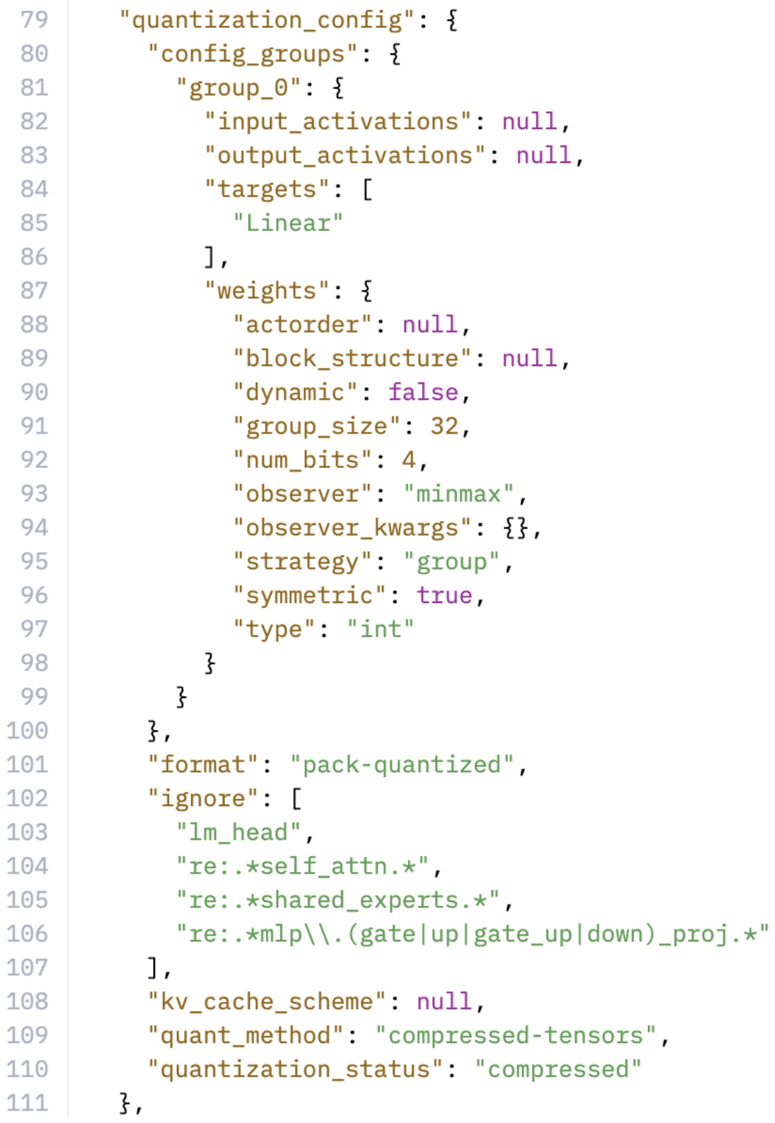


moonshotai / **Kimi-K2-Thinking**

Kimi-K2-Thinking 594 GB

bigmoyan fix-default-system-prompt (#12) 6126819 VERIFIED

docs		
figures		
.gitattributes	Safe	1.85 kB
LICENSE	Safe	1.46 kB
README.md	Safe	16.2 kB
THIRD_PARTY_NOTICES.md	Safe	1.67 kB
chat_template.jinja	Safe	3.45 kB
config.json	Safe	3.83 kB



```
  "quantization_config": {
    "config_groups": {
      "group_0": {
        "input_activations": null,
        "output_activations": null,
        "targets": [
          "Linear"
        ],
        "weights": {
          "actorder": null,
          "block_structure": null,
          "dynamic": false,
          "group_size": 32,
          "num_bits": 4,
          "observer": "minmax",
          "observer_kwargs": {},
          "strategy": "group",
          "symmetric": true,
          "type": "int"
        }
      }
    },
    "format": "pack-quantized",
    "ignore": [
      "lm_head",
      "re:.*self_attn.*",
      "re:.*shared_experts.*",
      "re:.*mlp\\.(gate|up|gate_up|down)_proj.*"
    ],
    "kv_cache_scheme": null,
    "quant_method": "compressed-tensors",
    "quantization_status": "compressed"
  },
```

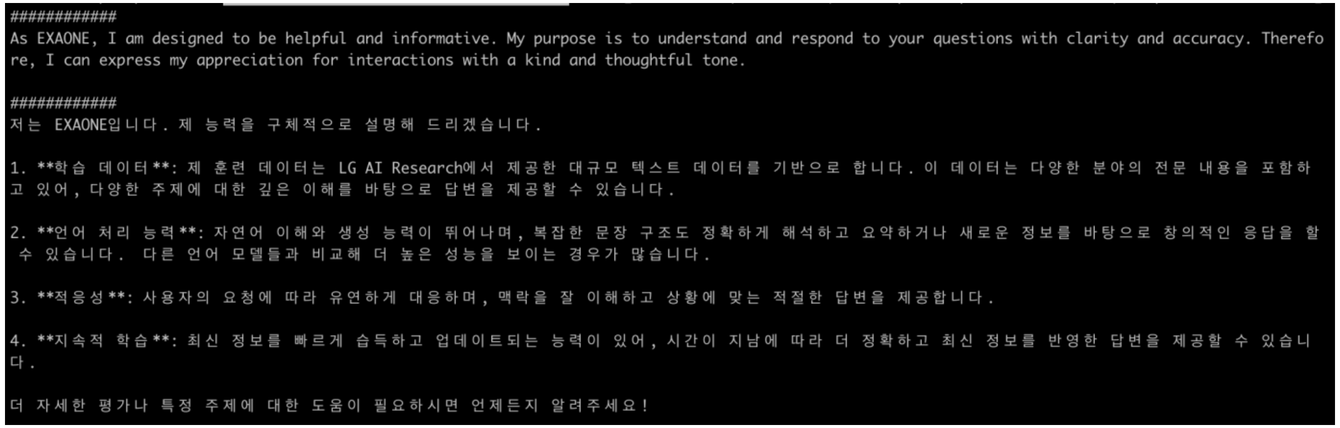
https://huggingface.co/moonshotai/Kimi-K2-Thinking/blob/main/config.json#L79-L111

추론 엔진 적용 - vLLM

- vLLM 은 기본적으로 여러 Quantization 모델의 추론을 지원
- 지원하지 않는 Quantization 기법의 경우 vLLM 에서 동작할 수 있도록 코드 구현 필요

```python
from vllm import LLM, SamplingParams

prompts = [
    [{"role": "user", "content": "Explain how wonderful you are"}],
    [{"role": "user", "content": "너가 얼마나 대단한지 설명해 봐"}],
]
sampling_params = SamplingParams(temperature=0.0, top_p=1.0, max_tokens=256)

llm = LLM(model="EXAONE-4.0-1.2B-GPTQ")

outputs = llm.chat(prompts, sampling_params)

for output in outputs:
    print("############")
    print(output.outputs[0].text)
    print()
```

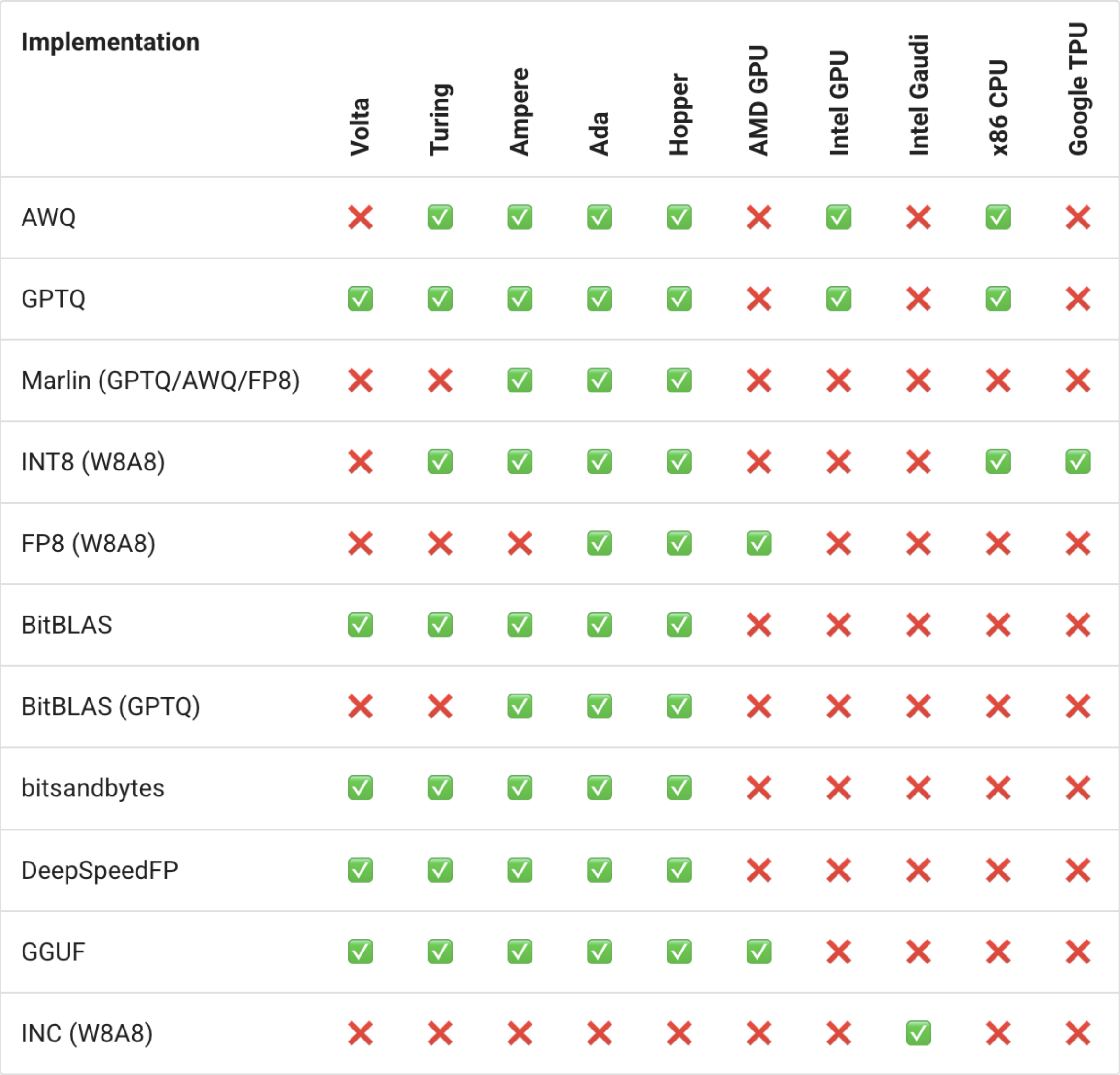
```
python3 vllm_inference.py
```



```
############
As EXAONE, I am designed to be helpful and informative. My purpose is to understand and respond to your questions with clarity and accuracy. Therefo
re, I can express my appreciation for interactions with a kind and thoughtful tone.

############
저는 EXAONE입니다. 제 능력을 구체적으로 설명해 드리겠습니다.

1. **학습 데이터**: 제 훈련 데이터는 LG AI Research에서 제공한 대규모 텍스트 데이터를 기반으로 합니다. 이 데이터는 다양한 분야의 전문 내용을 포함하
고 있어, 다양한 주제에 대한 깊은 이해를 바탕으로 답변을 제공할 수 있습니다.

2. **언어 처리 능력**: 자연어 이해와 생성 능력이 뛰어나며, 복잡한 문장 구조도 정확하게 해석하고 요약하거나 새로운 정보를 바탕으로 창의적인 응답을 할
 수 있습니다. 다른 언어 모델들과 비교해 더 높은 성능을 보이는 경우가 많습니다.

3. **적응성**: 사용자의 요청에 따라 유연하게 대응하며, 맥락을 잘 이해하고 상황에 맞는 적절한 답변을 제공합니다.

4. **지속적 학습**: 최신 정보를 빠르게 습득하고 업데이트되는 능력이 있어, 시간이 지남에 따라 더 정확하고 최신 정보를 반영한 답변을 제공할 수 있습니
다.

더 자세한 평가나 특정 주제에 대한 도움이 필요하시면 언제든지 알려주세요!
```

vLLM quickstart : https://docs.vllm.ai/en/latest/getting_started/quickstart

추론 엔진 적용 - vLLM



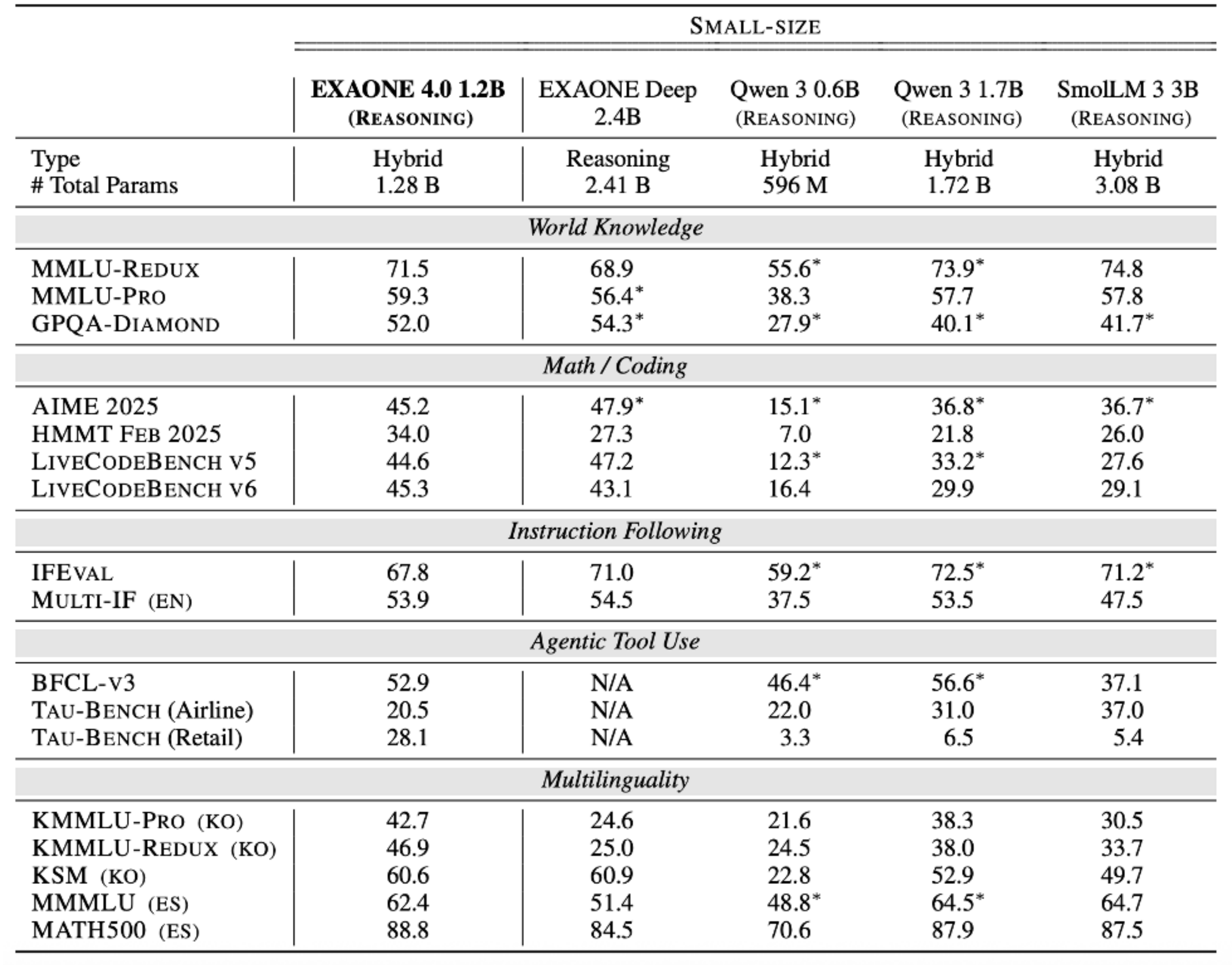
Implementation	Volta	Turing	Ampere	Ada	Hopper	AMD GPU	Intel GPU	Intel Gaudi	x86 CPU	Google TPU
AWQ	❌	✅	✅	✅	✅	❌	✅	❌	✅	❌
GPTQ	✅	✅	✅	✅	✅	❌	✅	❌	✅	❌
Marlin (GPTQ/AWQ/FP8)	❌	❌	✅	✅	✅	❌	❌	❌	❌	❌
INT8 (W8A8)	❌	✅	✅	✅	✅	❌	❌	❌	✅	✅
FP8 (W8A8)	❌	❌	❌	✅	✅	✅	❌	❌	❌	❌
BitBLAS	✅	✅	✅	✅	✅	❌	❌	❌	❌	❌
BitBLAS (GPTQ)	❌	❌	✅	✅	✅	❌	❌	❌	❌	❌
bitsandbytes	✅	✅	✅	✅	✅	❌	❌	❌	❌	❌
DeepSpeedFP	✅	✅	✅	✅	✅	❌	❌	❌	❌	❌
GGUF	✅	✅	✅	✅	✅	✅	❌	❌	❌	❌
INC (W8A8)	❌	❌	❌	❌	❌	❌	❌	✅	❌	❌

GPU 아키텍쳐 별 GPU 종류

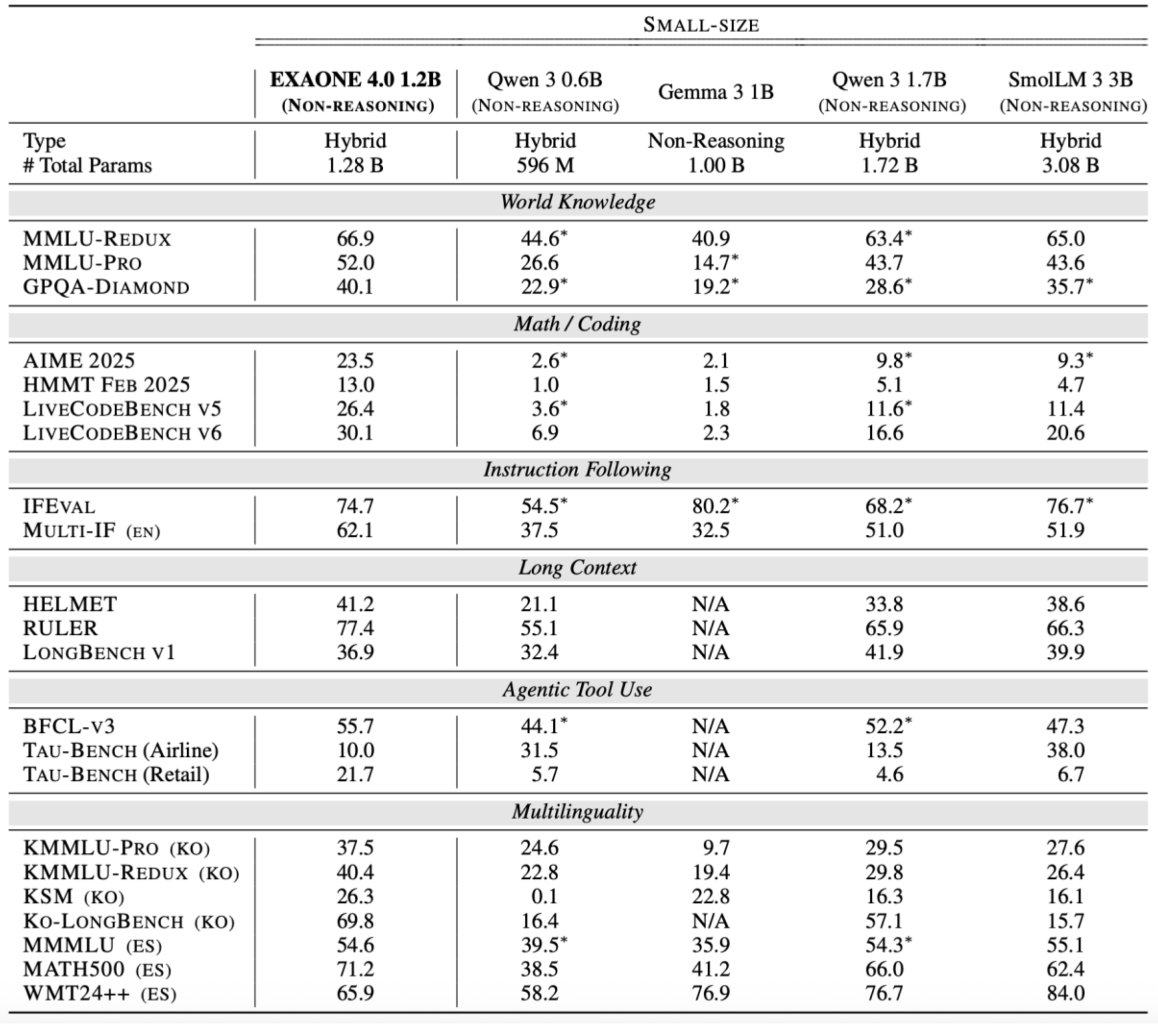
- Volta : V100 …
- Turing : T4, GeForce RTX 20 시리즈 …
- Ampere : A100, A10 …
- Ada Lovelace : GeForce RTX 40 시리즈, L4 …
- Hopper : H100, H800 …

https://github.com/vllm-project/vllm/tree/main/docs/features/quantization

평가 지표 - Accuracy



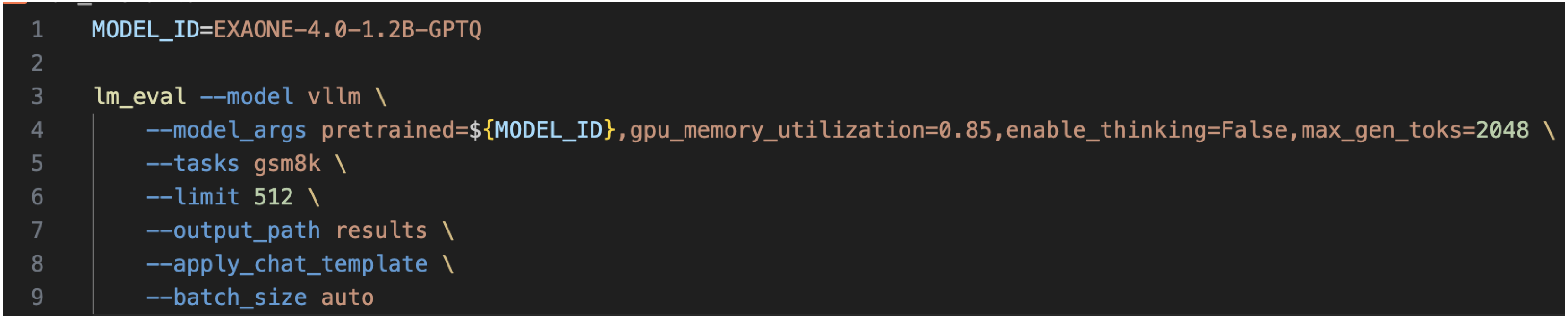
	Small-size				
	EXAONE 4.0 1.2B **(Reasoning)**	EXAONE Deep 2.4B	Qwen 3 0.6B (Reasoning)	Qwen 3 1.7B (Reasoning)	SmolLM 3 3B (Reasoning)
Type # Total Params	Hybrid 1.28 B	Reasoning 2.41 B	Hybrid 596 M	Hybrid 1.72 B	Hybrid 3.08 B
World Knowledge					
MMLU-Redux	71.5	68.9	55.6*	73.9*	74.8
MMLU-Pro	59.3	56.4*	38.3	57.7	57.8
GPQA-Diamond	52.0	54.3*	27.9*	40.1*	41.7*
Math / Coding					
AIME 2025	45.2	47.9*	15.1*	36.8*	36.7*
HMMT Feb 2025	34.0	27.3	7.0	21.8	26.0
LiveCodeBench v5	44.6	47.2	12.3*	33.2*	27.6
LiveCodeBench v6	45.3	43.1	16.4	29.9	29.1
Instruction Following					
IFEval	67.8	71.0	59.2*	72.5*	71.2*
Multi-IF (EN)	53.9	54.5	37.5	53.5	47.5
Agentic Tool Use					
BFCL-v3	52.9	N/A	46.4*	56.6*	37.1
Tau-Bench (Airline)	20.5	N/A	22.0	31.0	37.0
Tau-Bench (Retail)	28.1	N/A	3.3	6.5	5.4
Multilinguality					
KMMLU-Pro (KO)	42.7	24.6	21.6	38.3	30.5
KMMLU-Redux (KO)	46.9	25.0	24.5	38.0	33.7
KSM (KO)	60.6	60.9	22.8	52.9	49.7
MMMLU (ES)	62.4	51.4	48.8*	64.5*	64.7
MATH500 (ES)	88.8	84.5	70.6	87.9	87.5



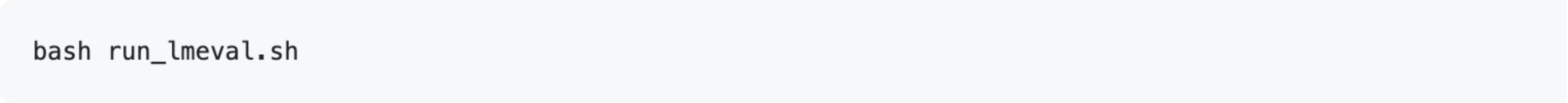
	Small-size				
	EXAONE 4.0 1.2B **(Non-reasoning)**	Qwen 3 0.6B (Non-reasoning)	Gemma 3 1B	Qwen 3 1.7B (Non-reasoning)	SmolLM 3 3B (Non-reasoning)
Type # Total Params	Hybrid 1.28 B	Hybrid 596 M	Non-Reasoning 1.00 B	Hybrid 1.72 B	Hybrid 3.08 B
World Knowledge					
MMLU-Redux	66.9	44.6*	40.9	63.4*	65.0
MMLU-Pro	52.0	26.6	14.7*	43.7	43.6
GPQA-Diamond	40.1	22.9*	19.2*	28.6*	35.7*
Math / Coding					
AIME 2025	23.5	2.6*	2.1	9.8*	9.3*
HMMT Feb 2025	13.0	1.0	1.5	5.1	4.7
LiveCodeBench v5	26.4	3.6*	1.8	11.6*	11.4
LiveCodeBench v6	30.1	6.9	2.3	16.6	20.6
Instruction Following					
IFEval	74.7	54.5*	80.2*	68.2*	76.7*
Multi-IF (EN)	62.1	37.5	32.5	51.0	51.9
Long Context					
HELMET	41.2	21.1	N/A	33.8	38.6
RULER	77.4	55.1	N/A	65.9	66.3
LongBench v1	36.9	32.4	N/A	41.9	39.9
Agentic Tool Use					
BFCL-v3	55.7	44.1*	N/A	52.2*	47.3
Tau-Bench (Airline)	10.0	31.5	N/A	13.5	38.0
Tau-Bench (Retail)	21.7	5.7	N/A	4.6	6.7
Multilinguality					
KMMLU-Pro (KO)	37.5	24.6	9.7	29.5	27.6
KMMLU-Redux (KO)	40.4	22.8	19.4	29.8	26.4
KSM (KO)	26.3	0.1	22.8	16.3	16.1
Ko-LongBench (KO)	69.8	16.4	N/A	57.1	15.7
MMMLU (ES)	54.6	39.5*	35.9	54.3*	55.1
MATH500 (ES)	71.2	38.5	41.2	66.0	62.4
WMT24++ (ES)	65.9	58.2	76.9	76.7	84.0

평가 지표 - Accuracy

- 오픈소스 평가 프레임워크인 lm-evaluation-harness 를 이용하여 평가 진행
- gsm8k외에도 평가할 수 있는 많은 태스크가 존재

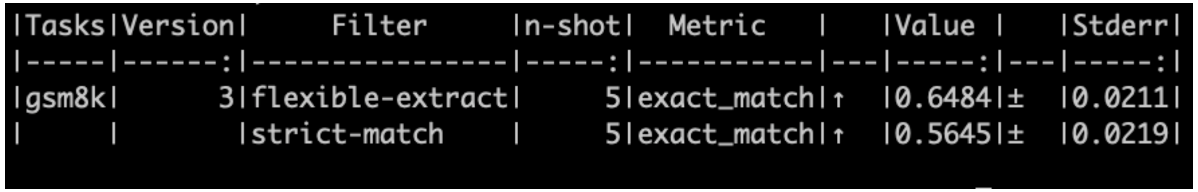


```bash
MODEL_ID=EXAONE-4.0-1.2B-GPTQ

lm_eval --model vllm \
    --model_args pretrained=${MODEL_ID},gpu_memory_utilization=0.85,enable_thinking=False,max_gen_toks=2048 \
    --tasks gsm8k \
    --limit 512 \
    --output_path results \
    --apply_chat_template \
    --batch_size auto
```



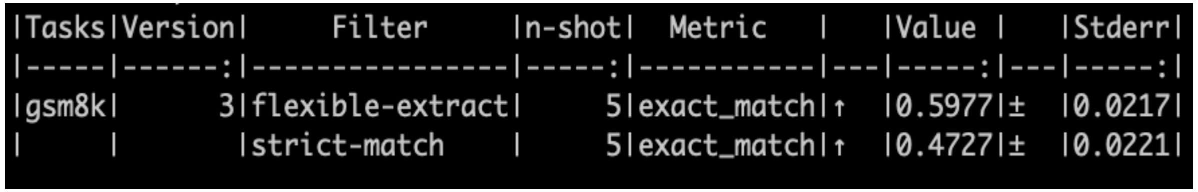
```
bash run_lmeval.sh
```

EXAONE-4.0-1.2B



```
|Tasks|Version|     Filter     |n-shot|  Metric   |   |Value |   |Stderr|
|-----|------:|----------------|-----:|-----------|---|-----:|---|-----:|
|gsm8k|      3|flexible-extract|     5|exact_match|↑  |0.6484|±  |0.0211|
|     |       |strict-match    |     5|exact_match|↑  |0.5645|±  |0.0219|
```

Quantized EXAONE-4.0-1.2B

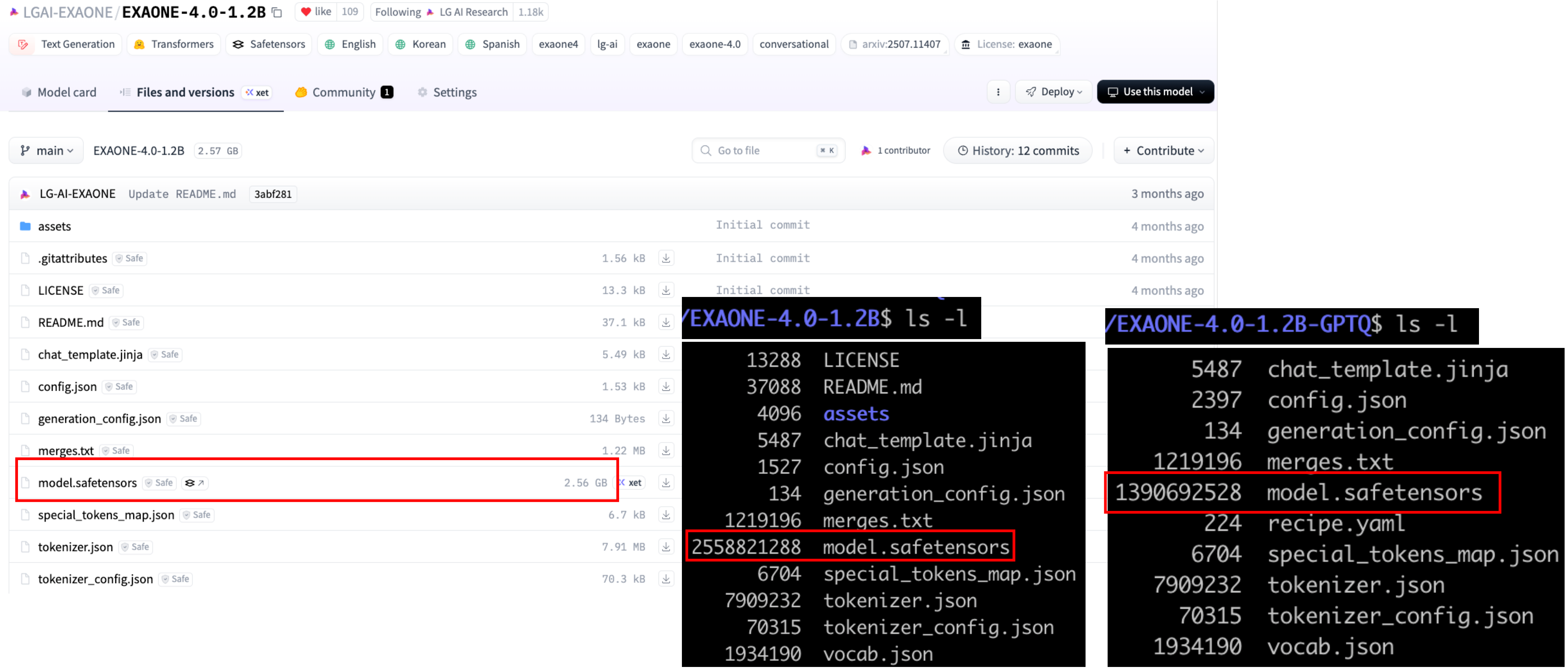


```
|Tasks|Version|     Filter     |n-shot|  Metric   |   |Value |   |Stderr|
|-----|------:|----------------|-----:|-----------|---|-----:|---|-----:|
|gsm8k|      3|flexible-extract|     5|exact_match|↑  |0.5977|±  |0.0217|
|     |       |strict-match    |     5|exact_match|↑  |0.4727|±  |0.0221|
```

https://github.com/EleutherAI/lm-evaluation-harness/tree/main/lm_eval/tasks#tasks

평가 지표 - Memory

- 최종적으로 저장되는 safetensors 파일의 크기를 측정



LGAI-EXAONE / **EXAONE-4.0-1.2B** like 109 Following LG AI Research 1.18k

Text Generation · Transformers · Safetensors · English · Korean · Spanish · exaone4 · lg-ai · exaone · exaone-4.0 · conversational · arxiv:2507.11407 · License: exaone

main EXAONE-4.0-1.2B 2.57 GB Go to file 1 contributor History: 12 commits + Contribute

LG-AI-EXAONE Update README.md 3abf281 3 months ago

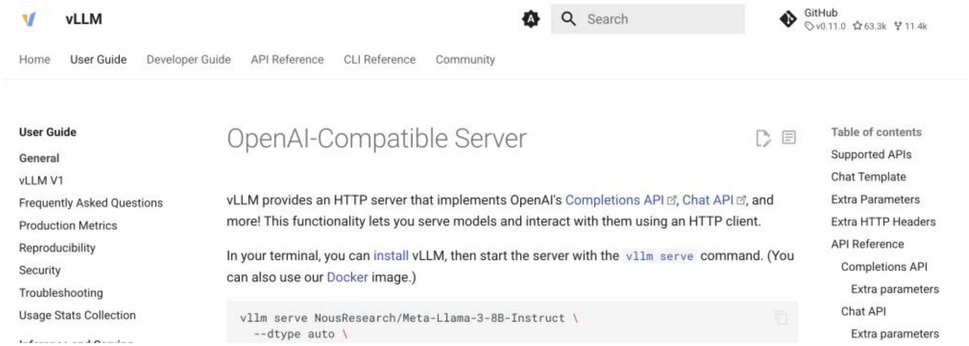
assets		Initial commit	4 months ago
.gitattributes Safe	1.56 kB	Initial commit	4 months ago
LICENSE Safe	13.3 kB	Initial commit	4 months ago
README.md Safe	37.1 kB		
chat_template.jinja Safe	5.49 kB		
config.json Safe	1.53 kB		
generation_config.json Safe	134 Bytes		
merges.txt Safe	1.22 MB		
model.safetensors Safe	2.56 GB xet		
special_tokens_map.json Safe	6.7 kB		
tokenizer.json Safe	7.91 MB		
tokenizer_config.json Safe	70.3 kB		

```
/EXAONE-4.0-1.2B$ ls -l
     13288  LICENSE
     37088  README.md
      4096  assets
      5487  chat_template.jinja
      1527  config.json
       134  generation_config.json
   1219196  merges.txt
2558821288  model.safetensors
      6704  special_tokens_map.json
   7909232  tokenizer.json
     70315  tokenizer_config.json
   1934190  vocab.json
```

```
/EXAONE-4.0-1.2B-GPTQ$ ls -l
      5487  chat_template.jinja
      2397  config.json
       134  generation_config.json
   1219196  merges.txt
1390692528  model.safetensors
       224  recipe.yaml
      6704  special_tokens_map.json
   7909232  tokenizer.json
     70315  tokenizer_config.json
   1934190  vocab.json
```

부록 - OpenAI Compatible

- 최근에는 vLLM과 같은 추론엔진을 OpenAI Compatible Server 형태로 구동하고 평가 프레임워크에서 API 를 호출하는 형태의 평가 방식이 인기있음
- 개발자들 사이에서 OpenAI 라이브러리가 대중화되고 어떤 평가든 일관된 포맷으로 평가가 가능해 쉽게 구현 및 구동이 쉽다는 장점이 있음
- OpenAI Compatible은 오픈소스계에서 최소 조건이 되어가고 있음
- 추론엔진은 OpenAI Compatible Server를 제공하고 평가 프레임워크들은 OpenAI Compatible endpoint를 사용함

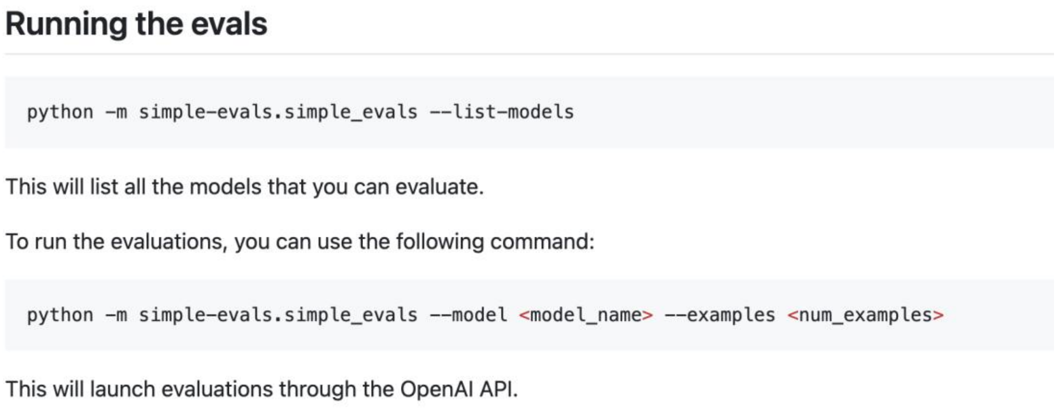


OpenAI-Compatible Server

vLLM provides an HTTP server that implements OpenAI's Completions API, Chat API, and more! This functionality lets you serve models and interact with them using an HTTP client.

In your terminal, you can install vLLM, then start the server with the `vllm serve` command. (You can also use our Docker image.)

```
vllm serve NousResearch/Meta-Llama-3-8B-Instruct \
  --dtype auto \
```



Running the evals

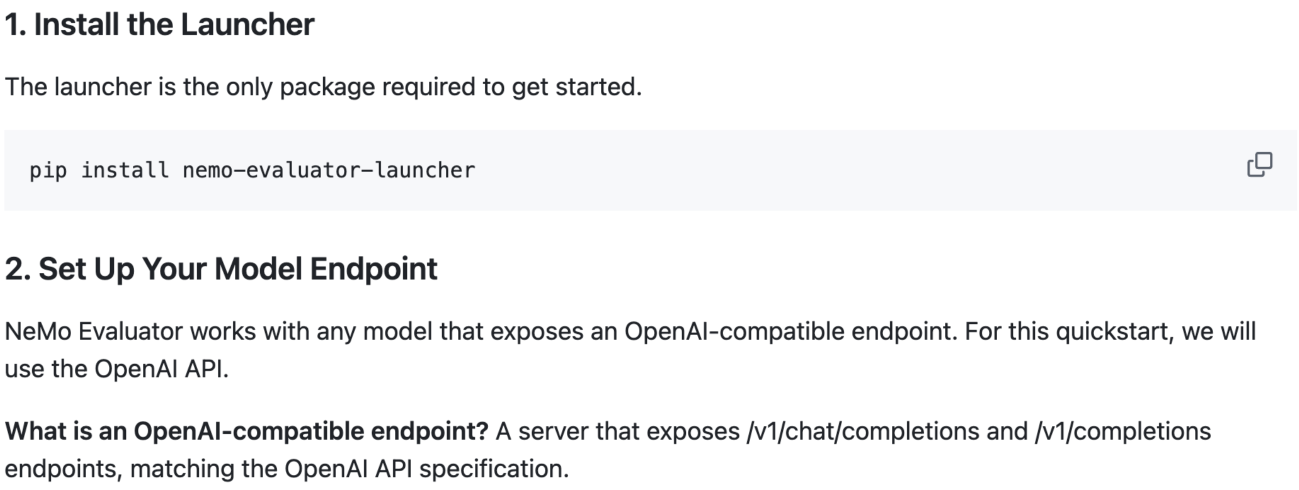
```
python -m simple-evals.simple_evals --list-models
```

This will list all the models that you can evaluate.

To run the evaluations, you can use the following command:

```
python -m simple-evals.simple_evals --model <model_name> --examples <num_examples>
```

This will launch evaluations through the OpenAI API.



1. Install the Launcher

The launcher is the only package required to get started.

```
pip install nemo-evaluator-launcher
```

2. Set Up Your Model Endpoint

NeMo Evaluator works with any model that exposes an OpenAI-compatible endpoint. For this quickstart, we will use the OpenAI API.

What is an OpenAI-compatible endpoint? A server that exposes /v1/chat/completions and /v1/completions endpoints, matching the OpenAI API specification.

https://docs.vllm.ai/en/latest/serving/openai_compatible_server/

https://github.com/openai/simple-evals

https://github.com/NVIDIA-NeMo/Evaluator