

LLM Serving Methods Evaluation

Name: Sandaruwan P.K.U.I.

Code: [code 1](#), [code 2](#)

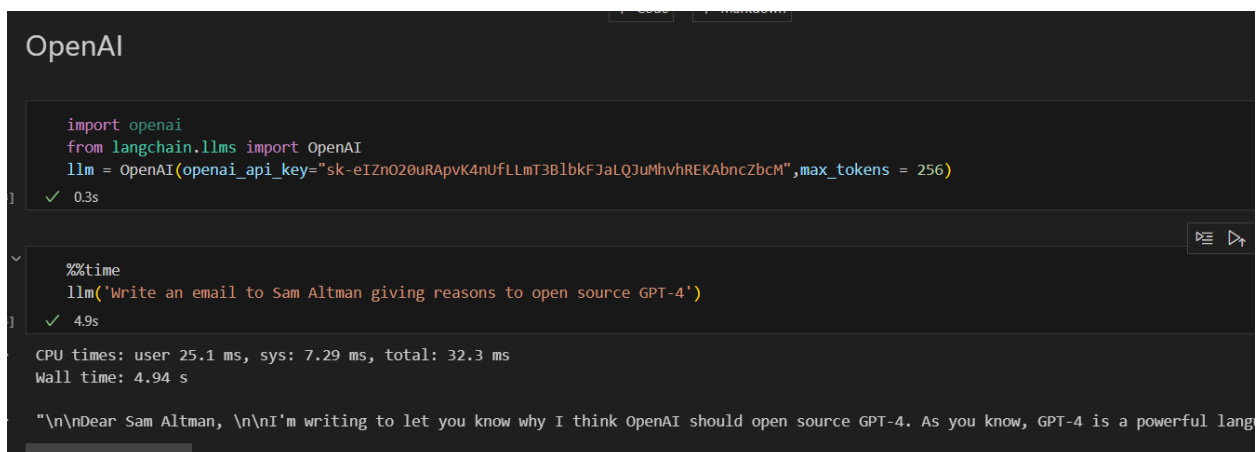
This document provides a comprehensive examination of inference in text generation and vLLM projects. It also includes evaluations of techniques such as Quantization, Pruning, and Knowledge Distillation aimed at enhancing the performance of Language Model (LLM) models in resource-constrained environments.

1. Directly Using Huggingface models
2. Huggingface models with Text Generation Inference (TGI)
3. Huggingface models with vLLM

g4.dn.xlarge AWS EC2 instance was used in these evaluations unless specifically said otherwise. This instance uses an Intel Xeon processor with 4 virtual CPUs, operating at 2.5GHz, 16GB of memory, and a ***single Nvidia T4 GPU*** with 16GB of dedicated memory. Nvidia T4 GPU Compute Capability is 7.5

All of the evaluations are based on a max token limit of 256 for output response. Same prompt has been used.

OpenAI



```
OpenAI

import openai
from langchain.llms import OpenAI
llm = OpenAI(openai_api_key="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",max_tokens = 256)

✓ 0.3s

%%time
llm('Write an email to Sam Altman giving reasons to open source GPT-4')

✓ 4.9s

CPU times: user 25.1 ms, sys: 7.29 ms, total: 32.3 ms
Wall time: 4.94 s

"\n\nDear Sam Altman, \n\nI'm writing to let you know why I think OpenAI should open source GPT-4. As you know, GPT-4 is a powerful lang
```

Remarks:

Response is relevant to the question. Replied within 4.94 seconds.

1. Directly Using HuggingfaceHub models

Downloading the model and using it with the pipeline:

Model name: **meta-llama/Llama-2-7b-chat-hf**

```
%%time
prompt = 'Write an email to Sam Altman giving reasons to open source GPT-4'
get_llama_response(prompt)
✓ 14.0s
```

Chatbot: Write an email to Sam Altman giving reasons to open source GPT-4.

Subject: Open Sourcing GPT-4: Why it's a Good Idea

Dear Sam Altman,

I hope this email finds you well. I am writing to you today to express my belief that GPT-4, the latest language model developed by OpenAI, is a significant milestone in the field of artificial intelligence. First and foremost, open sourcing GPT-4 would allow for greater transparency and accountability in the development and deployment of such powerful models. This would enable researchers and developers from around the world to study, improve, and build upon the model, potentially leading to faster innovation and more robust, safe AI systems. Secondly, open sourcing GPT-4 would facilitate collaboration and innovation in the field of natural language processing. By making the model's architecture and training data (where possible) available, the community can identify areas for improvement and develop new applications that OpenAI might not have considered. Finally, open sourcing GPT-4 would help democratize access to state-of-the-art AI technology. Researchers and developers who lack the resources of a large tech company could still benefit from the capabilities of a model like GPT-4, accelerating their work and contributing to the broader AI ecosystem.

CPU times: user 14.1 s, sys: 0 ns, total: 14.1 s
Wall time: 14.1 s

Model name: **mistralai/Mistral-7B-v0.1**

```
%%time
prompt = 'Write an email to Sam Altman giving reasons to open source GPT-4'
get_llama_response(prompt)
✓ 1m 41.4s
```

Setting `pad_token_id` to `eos_token_id`:2 for open-end generation.

Chatbot: Write an email to Sam Altman giving reasons to open source GPT-4 and GPT-3.

I'd start with "Dear Mr. Altman", and I'd end my email with "Sincerely," and I'd make sure to keep my email short and to the point.

I would tell you that I was really excited to be writing to you about the possibility of opening up the GPT-4 and GPT-3 algorithms to the public. I believe that open sourcing these models would greatly benefit the AI community and the world at large.

I'd also tell you that I thought it would be great to have an AI that could understand and write in multiple languages, and I also thought it would be great to have an AI that could help with a wide variety of tasks.

I'd then tell you that I'd been reading about the work you've been doing with these models, and I thought it would be really great to see the results of your research.

I'd then say that I was very interested in helping out with the project, and I was looking forward to learning more about the models and how they were being used.

I'd also ask if I could come visit OpenAI to see the models in action and learn more about the team.

CPU times: user 1min 40s, sys: 1.19 s, total: 1min 41s
Wall time: 1min 41s

Remarks:

Using LLMs with pipeline takes up a lot of CPU's Memory. These 7B parameter models took up to 16GB of Memory during the inference. Response quality is good. But it took some time to answer. Llama2 model responded with in 14.1seconds. But Mistral 7B took more than 1 minute. Both are slower compared to OpenAI

2. HuggingfaceHub models with Text Generation Inference (TGI)

To run the Text Generation Inference, we need a GPU which has at least a compute capability of 8.0. So, the instance I am using is not enough. I will use [runpod](#) servers to use TGI.

Model Name: TheBloke/Llama-2-7b-chat-fp16

```
%%time
prompt = generate_prompt(
    "Write an email to Sam Altman giving reasons to open source GPT-4",
    system_prompt=DWIGHT_SYSTEM_PROMPT,
)
response = make_request(prompt)
✓ 8.2s

CPU times: total: 141 ms
Wall time: 8.19 s

print(response.json()["generated_text"].strip())
✓ 0.0s

Subject: Open Sourcing GPT-4: A Beneficial Move for the AI Community

Dear Sam Altman,

I hope this email finds you well. As a fellow AI enthusiast and advocate for open-source technology, I am writing to express my strong support for the idea of open-sourcing GPT-4. Here are some reasons why I think open-sourcing GPT-4 is a good idea:
1. Promotes Collaboration: By making GPT-4 open-source, you will be fostering collaboration among researchers, developers, and organizations. This will lead to a faster pace of in
2. Increases Transparency: Open-sourcing GPT-4 will provide a deeper understanding of how the model works, its limitations, and potential biases. This transparency is crucial in a
3. Enhances Reliability: With GPT-4 being open
```

Remarks:

The NVIDIA T4, with a compute capability of 7.5, falls short of the minimum requirement of 8.0 for Text Generation Inference. As an alternative, I opted for the NVIDIA RTX A4500, boasting 1 GPU, 20GB VRAM, 62GB RAM, and 12 vCPU, which yielded a response time of 8.19 seconds. Although I utilized it via an API key from runpod, connecting directly to runpod servers is expected to deliver an even faster response. Notably, Text Generation Inference using the RTX A4500 outperformed direct usage of models from the Hugging Face Hub in terms of speed.

3. Huggingface models with vLLM

Model name : TheBloke/Mistral-7B-Instruct-v0.1-AWQ

```
%%time
prompt = 'Write an email to Sam Altman giving reasons to open source GPT-4'
generated_text = generate(prompt)
parse_text(generated_text)
✓ 7.3s
```

Processed prompts: 100%|██████████| 1/1 [00:07<00:00, 7.39s/it]

Dear Sam Altman, I hope this email finds you well. As an avid user of OpenAI's latest model, GPT-4, I wanted to take a moment to express my gratitude for the incredible work that has gone into its development. I have been amazed by the model's ability to understand and generate complex language, and I believe that it has tremendous potential to be a game-changer in a wide range of fields. With that in mind, I would like to suggest that OpenAI consider releasing GPT-4 as open source. There are several reasons why I believe this would be a wise decision: 1. Increased collaboration: By making GPT-4 open source, developers and researchers from around the world would have access to the model's source code, allowing them to contribute to its development and build upon its capabilities. This could lead to a wealth of new and innovative applications for the model, as well as improvements in its accuracy and performance. 2. Greater transparency: Open source software is often more transparent than proprietary software, as the source code is publicly available for anyone to view. This can help build trust in the software, as users

CPU times: user 7.36 s, sys: 16.1 ms, total: 7.37 s
Wall time: 7.39 s

Remarks:

The system requirements for vLLM include the following:

- Operating System: Linux
- Python Version: 3.8 – 3.11
- GPU: Compute capability of 7.0 or higher, such as V100, T4, RTX20xx, A100, L4, H100, etc.

Despite not demanding high-end GPUs, vLLM stands out as the fastest inference option for Open Source LLMs. There is no degradation in response quality, and it delivered the response within 7.39 seconds. Running a 7B LLM only requires 16GB VRAM, proving to be more than sufficient for optimal performance.

Using vLLM with Langchain

```
from langchain.llms import VLLM
```

✓ 0.2s

```
%%time
```

```
llm = VLLM(  
    model="TheBloke/Mistral-7B-Instruct-v0.1-AWQ",  
    quantization='awq',  
    trust_remote_code=True,  
    dtype = 'half',  
    temperature=0.8,  
    gpu_memory_utilization=.95,  
    max_model_len= 2000,  
    max_split_size_mb = 500,  
)
```

✗ 8.4s

```
93     download_dir=values["download_dir"],
```

```
... --> 156 out = ops.awq_gemm(reshaped_x, qweight, scales, qzeros, pack_factor)
```

```
157 if bias is not None:
```

```
158     out = out + bias
```

OutOfMemoryError: CUDA out of memory. Tried to allocate 14.00 GiB. GPU 0 has a total capacity of 14.58 GiB of which 9.43 GiB is free. Inc
Output is truncated. View as a [scrollable element](#) or open in a [text editor](#). Adjust cell output [settings](#)...

+ Code

+ Markdown

```
%%time
```

```
prompt = 'Write an email to Sam Altman giving reasons to open source GPT-4'
```

```
generated_text = generate(prompt)
```

```
parse_text(generated_text)
```

Remarks:

I encountered an issue when attempting to integrate vLLM with Langchain. While I had successfully used vLLM independently with a 16GB VRAM, integrating it with Langchain demanded a minimum of 26GB VRAM.

Quantization

GPTQ

```
%%time
outputs = pipe(
    prompt,
    max_new_tokens=256,
    do_sample=True,
    temperature=0.1,
    top_p=0.95
)
print(outputs[0]["generated_text"])

✓ 9.7s

/home/ubuntu/.local/lib/python3.10/site-packages/transformers/generation/utils.py:1547: UserWarning: You have modified
warnings.warn(
<|system|>
You are a friendly chatbot.</s>
<|user|>
Write an email to Sam Altman giving reasons to open source GPT-4|>
Subject: Request for Open Sourcing GPT-4

Dear Sam Altman,

I am writing to request the open sourcing of GPT-4, the latest and most advanced language model developed by OpenAI. As

Firstly, open sourcing GPT-4 would allow for greater collaboration and innovation in the field of AI. By making the mod

Secondly, open sourcing GPT-4 would promote greater transparency and accountability in the development of AI technology

CPU times: user 9.72 s, sys: 0 ns, total: 9.72 s
Wall time: 9.77 s
```

Remarks: GPTQ is a post-training quantization (PTQ) method for 4-bit quantization that focuses primarily on GPU inference and performance. The idea behind the method is that it will try to compress all weights to a 4-bit quantization by minimizing the mean squared error to that weight. During inference, it will dynamically dequantize its weights to float16 for improved performance whilst keeping memory low.

GGUF

```
%%time
outputs = pipe(prompt, max_new_tokens=256)
print(outputs[0]["generated_text"])

[21] ✓ 13.9s

... <|system|>
You are a friendly chatbot.</s>
<|user|>
Write an email to Sam Altman giving reasons to open source GPT-4|>
Subject: Request for Open Sourcing GPT-4

Dear Mr. Sam Altman,

I am writing to you today to request that OpenAI, the organization you lead, consider open sourcing the upcoming GPT-4 langua

Firstly, open sourcing GPT-4 will accelerate the pace of innovation in the field of natural language processing (NLP). The cu

Secondly, open sourcing GPT-4 will promote transparency and trust in the industry.

CPU times: user 37.6 s, sys: 278 ms, total: 37.9 s
Wall time: 14 s
```

Remarks: Although GPTQ does compression well, its focus on GPU can be a disadvantage if you do not have the hardware to run it. GGUF, previously GGML, is a quantization method that allows users to use the **CPU** to run an LLM but also offload some of its layers to the GPU for a speed up. Although using the CPU is generally slower than using a GPU for inference, it is an incredible format for those running models on CPU or Apple devices.

AWQ

```
%%time
output = llm.generate(prompt, sampling_params)
print(output[0].outputs[0].text)
✓ 7.3s
```

Processed prompts: 100%|██████████| 1/1 [00:07<00:00, 7.35s/it]
Subject: Request for Open Sourcing GPT-4

Dear Mr. Altman,

I am writing to request that OpenAI open sources its latest language model, GPT-4. As a technology enthusiast and a strong believer in

Firstly, open sourcing GPT-4 would enable a wider range of developers and researchers to build upon and contribute to the model's devel

Secondly, open sourcing GPT-4 would promote transparency and accountability in the development and deployment of AI technologies. This

Thirdly, open sourcing GPT-4 would foster collaboration and knowledge sharing within the AI community. This would help to accelerate th

CPU times: user 7.36 s, sys: 0 ns, total: 7.36 s
Wall time: 7.36 s

Remarks:

A new format on the block is [AWQ](#) (Activation-aware Weight Quantization) which is a quantization method similar to GPTQ. There are several differences between AWQ and GPTQ as methods but the most important one is that AWQ assumes that not all weights are equally important for an LLM's performance.

In other words, there is a small fraction of weights that will be skipped during quantization which helps with the quantization loss.