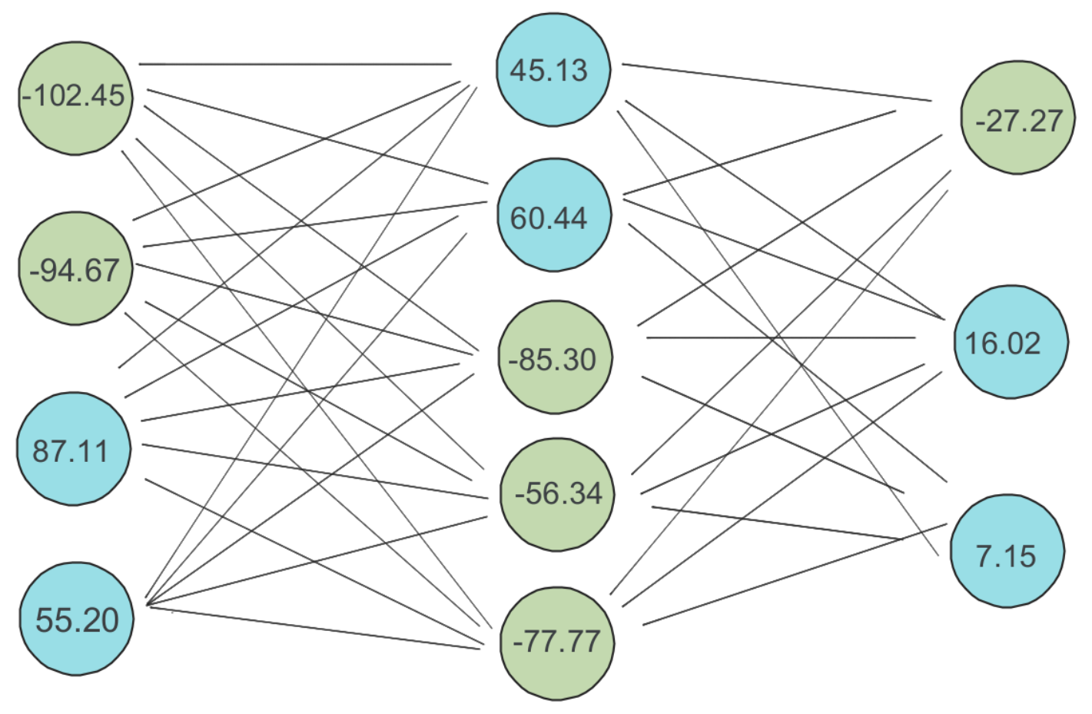
The Distributions of LLM's

Yosvel Baez, UNC Charlotte
Tyler Allen, College of Computing and Informatics

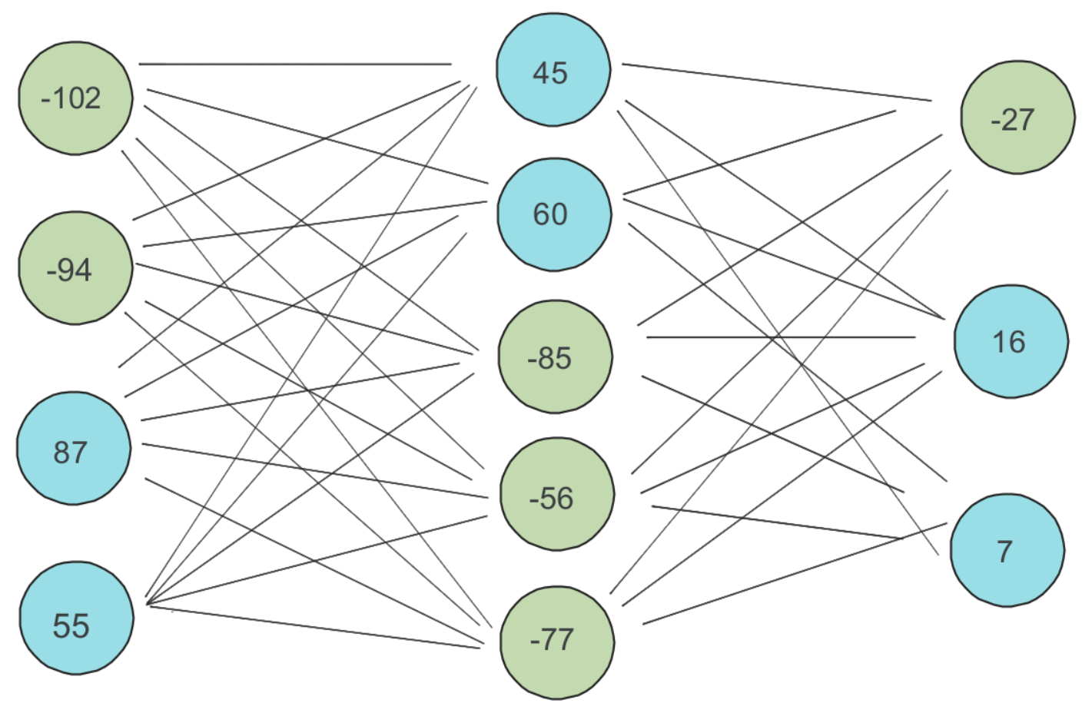
Introduction

Activation-aware Weight Quantization (AWQ), is a method for reducing the size and increasing the speed of LLMS by quantizing their weights which considers the activation patterns of the neural network. This approach enables the deployment of powerful AI models on devices with limited computational resources.



LLM's utilizing AWQ

- LLaMA
- OPT
- Vicuna



Distribution of AWQ ensures that LLM's are both memory and computation-efficient, enabling their deployment on devices with limited resources without sacrificing performance.

Methodology

Beginning Stage

- Familiarizing myself with command line language.
- Exploring GitHub and the repository chosen to conduct research on.

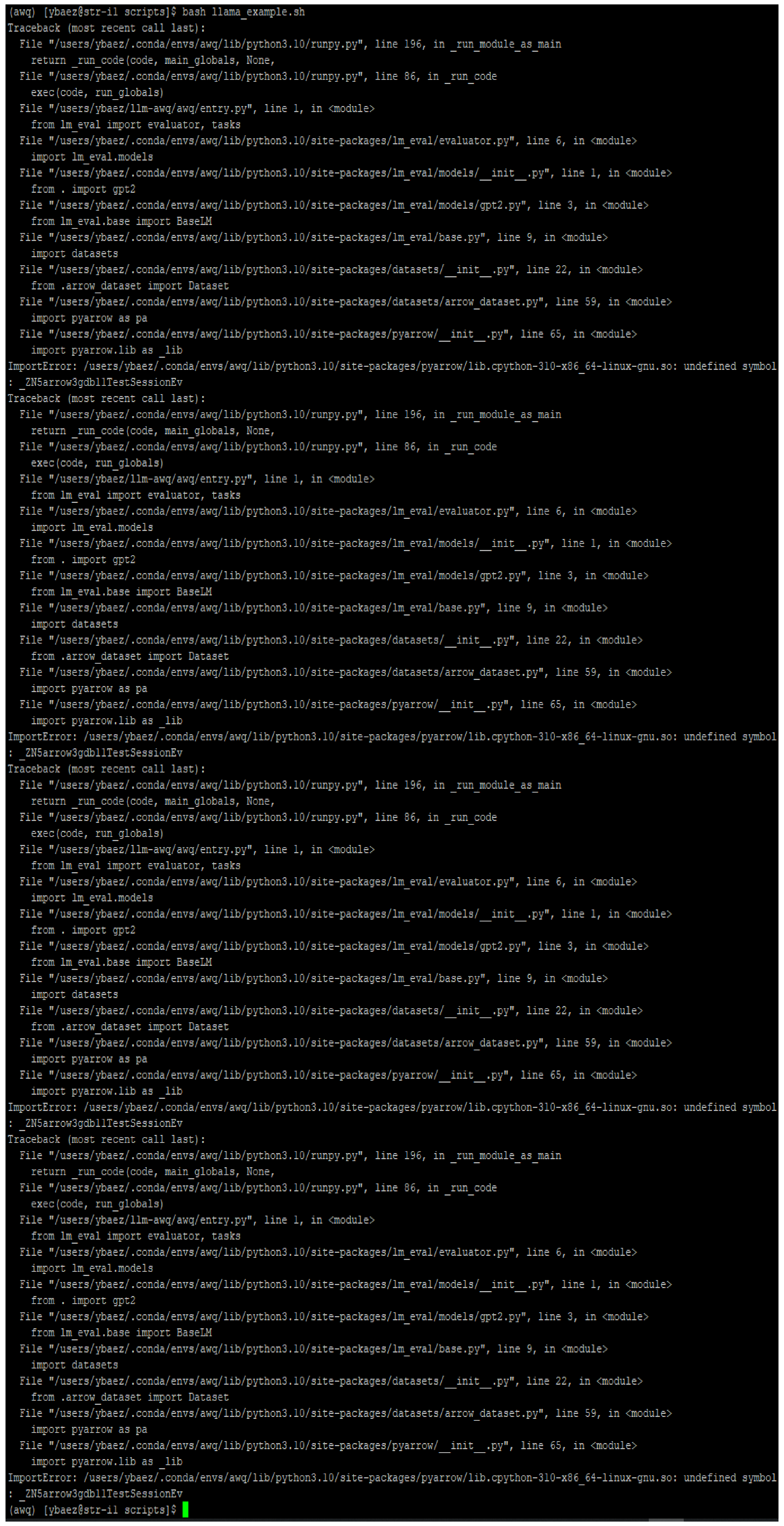
Mid-Stage

- Failed setting up my environment in VS Code through the instructions provided within the repository
- Moved onto the HPC server provided by UNCC due to complications from working on my own computer

Current Stage

- Ran all command line prompts presented in repository
- Waiting to be approved for an OPT data set from hugging face to complete my results.

Current Progress



```
(awq) [ybaez@str-il scripts]$ bash llama_example.sh
Traceback (most recent call last):
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/runpy.py", line 196, in _run_module_as_main
    return _run_code(code, main_globals, None,
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/runpy.py", line 86, in _run_code
    exec(code, run_globals)
  File "/users/ybaez/llm-awq/awq/entry.py", line 1, in <module>
    from lm_eval import evaluator, tasks
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/site-packages/lm_eval/evaluator.py", line 6, in <module>
    import lm_eval.models
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/site-packages/lm_eval/models/__init__.py", line 1, in <module>
    from . import gpt2
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/site-packages/lm_eval/models/gpt2.py", line 3, in <module>
    from lm_eval.base import BaseLM
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/site-packages/lm_eval/base.py", line 9, in <module>
    import datasets
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/site-packages/datasets/__init__.py", line 22, in <module>
    from .arrow_dataset import Dataset
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/site-packages/datasets/arrow_dataset.py", line 59, in <module>
    import pyarrow as pa
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/site-packages/pyarrow/__init__.py", line 65, in <module>
    import pyarrow.lib as _lib
ImportError: /users/ybaez/.conda/envs/awq/lib/python3.10/site-packages/pyarrow/lib.cpython-310-x86_64-linux-gnu.so: undefined symbol
: _ZN5arrow3gdb11TestSessionEv
Traceback (most recent call last):
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/runpy.py", line 196, in _run_module_as_main
    return _run_code(code, main_globals, None,
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/runpy.py", line 86, in _run_code
    exec(code, run_globals)
  File "/users/ybaez/llm-awq/awq/entry.py", line 1, in <module>
    from lm_eval import evaluator, tasks
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/site-packages/lm_eval/evaluator.py", line 6, in <module>
    import lm_eval.models
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/site-packages/lm_eval/models/__init__.py", line 1, in <module>
    from . import gpt2
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/site-packages/lm_eval/models/gpt2.py", line 3, in <module>
    from lm_eval.base import BaseLM
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/site-packages/lm_eval/base.py", line 9, in <module>
    import datasets
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/site-packages/datasets/__init__.py", line 22, in <module>
    from .arrow_dataset import Dataset
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/site-packages/datasets/arrow_dataset.py", line 59, in <module>
    import pyarrow as pa
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/site-packages/pyarrow/__init__.py", line 65, in <module>
    import pyarrow.lib as _lib
ImportError: /users/ybaez/.conda/envs/awq/lib/python3.10/site-packages/pyarrow/lib.cpython-310-x86_64-linux-gnu.so: undefined symbol
: _ZN5arrow3gdb11TestSessionEv
Traceback (most recent call last):
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/runpy.py", line 196, in _run_module_as_main
    return _run_code(code, main_globals, None,
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/runpy.py", line 86, in _run_code
    exec(code, run_globals)
  File "/users/ybaez/llm-awq/awq/entry.py", line 1, in <module>
    from lm_eval import evaluator, tasks
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/site-packages/lm_eval/evaluator.py", line 6, in <module>
    import lm_eval.models
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/site-packages/lm_eval/models/__init__.py", line 1, in <module>
    from . import gpt2
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/site-packages/lm_eval/models/gpt2.py", line 3, in <module>
    from lm_eval.base import BaseLM
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/site-packages/lm_eval/base.py", line 9, in <module>
    import datasets
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/site-packages/datasets/__init__.py", line 22, in <module>
    from .arrow_dataset import Dataset
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/site-packages/datasets/arrow_dataset.py", line 59, in <module>
    import pyarrow as pa
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/site-packages/pyarrow/__init__.py", line 65, in <module>
    import pyarrow.lib as _lib
ImportError: /users/ybaez/.conda/envs/awq/lib/python3.10/site-packages/pyarrow/lib.cpython-310-x86_64-linux-gnu.so: undefined symbol
: _ZN5arrow3gdb11TestSessionEv
Traceback (most recent call last):
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/runpy.py", line 196, in _run_module_as_main
    return _run_code(code, main_globals, None,
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/runpy.py", line 86, in _run_code
    exec(code, run_globals)
  File "/users/ybaez/llm-awq/awq/entry.py", line 1, in <module>
    from lm_eval import evaluator, tasks
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/site-packages/lm_eval/evaluator.py", line 6, in <module>
    import lm_eval.models
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/site-packages/lm_eval/models/__init__.py", line 1, in <module>
    from . import gpt2
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/site-packages/lm_eval/models/gpt2.py", line 3, in <module>
    from lm_eval.base import BaseLM
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/site-packages/lm_eval/base.py", line 9, in <module>
    import datasets
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/site-packages/datasets/__init__.py", line 22, in <module>
    from .arrow_dataset import Dataset
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/site-packages/datasets/arrow_dataset.py", line 59, in <module>
    import pyarrow as pa
  File "/users/ybaez/.conda/envs/awq/lib/python3.10/site-packages/pyarrow/__init__.py", line 65, in <module>
    import pyarrow.lib as _lib
ImportError: /users/ybaez/.conda/envs/awq/lib/python3.10/site-packages/pyarrow/lib.cpython-310-x86_64-linux-gnu.so: undefined symbol
: _ZN5arrow3gdb11TestSessionEv
(awq) [ybaez@str-il scripts]$
```

Screenshot of the work conducted through UNCC HPC server

Why utilize AWQ?

- To evaluate the efficacy of Activation-aware Weight Quantization (AWQ) in enhancing the execution efficiency of Large Language Models (LLMs) across diverse computational environments.
- To benchmark the execution performance of LLMs pre-AWQ and post-AWQ application, measuring improvements in speed, memory usage, and energy consumption on both standard and resource-constrained hardware platforms.

Motivation

- Advancing the field of machine learning by improving the efficiency of model deployment, especially for LLM's.
- Contributing to the development of more sustainable AI by reducing the computational and energy requirements for LLM's.
- Enabling the use of advanced AI models in real-world applications where resources are limited, such as mobile devices or other edge computing scenarios.

References

- MIT HAN LAB. (n.d.). tinychat. LLM-AWQ. Retrieved November 29, 2023, from https://github.com/mit-han-lab/llm-awq/tree/main/tinychat