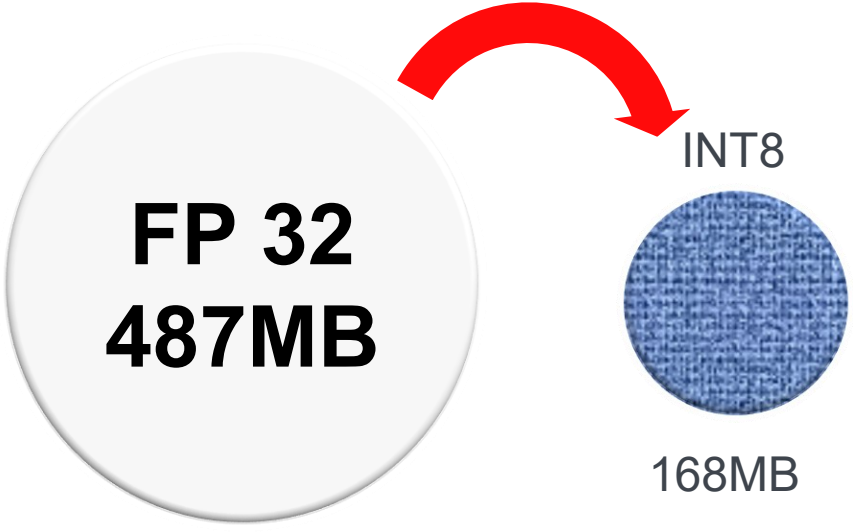
University of Stuttgart
Institute of Industrial Automation
and Software Engineering

Evaluation of Quantized Large Language Models for Semantic Interpretation and Reasoning within Industrial Automation Contexts

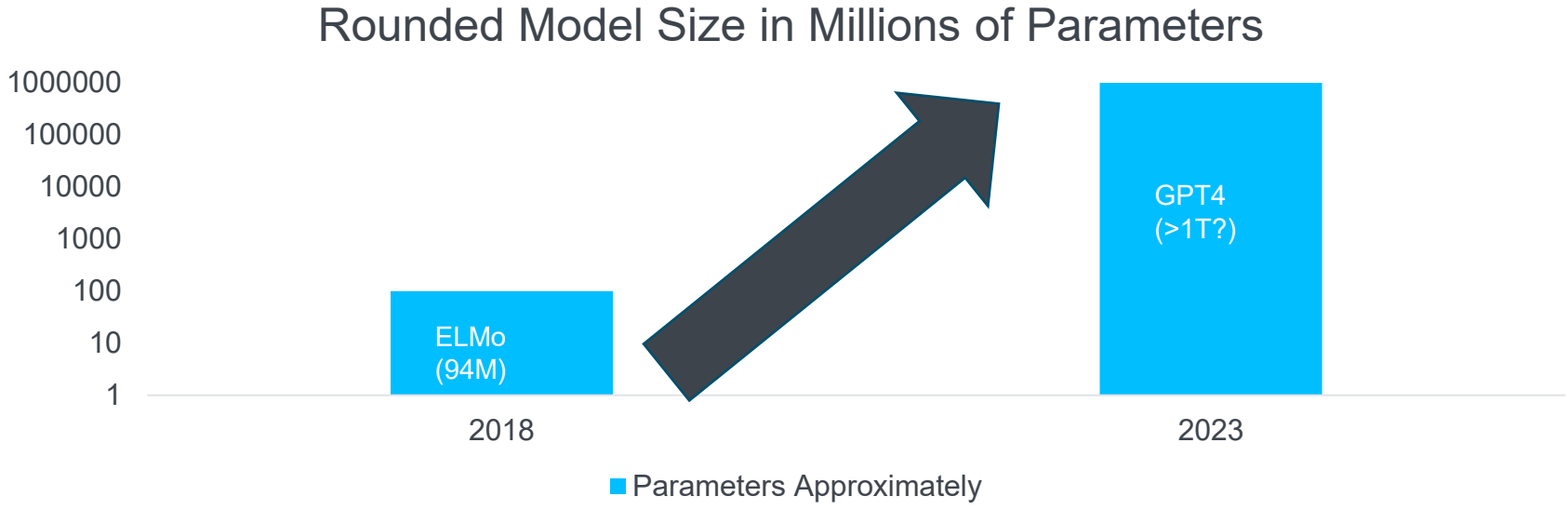
Belal Abulabn

Supervisor: Yuchen Xia

- **Motivation**
- **Basis**
- **Conception Design**
- **Implementation**
- **Evaluation and Verification**
- **Summary and Outlook**



Motivation

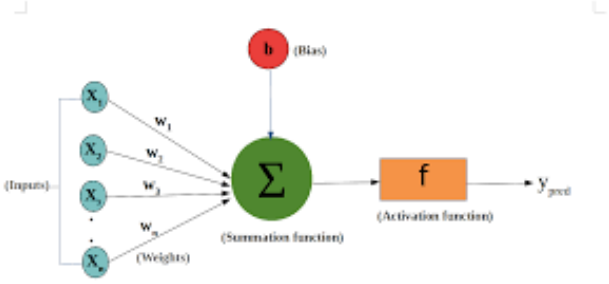


1. **High Computational Costs:** LLMs require significant computational power for training and inference, making them challenging to deploy in resource-constrained environments.

2. **Extensive Resource Requirements**: Beyond computation, LLMs demand considerable memory and storage, which are scarce on devices with limited processing capabilities.

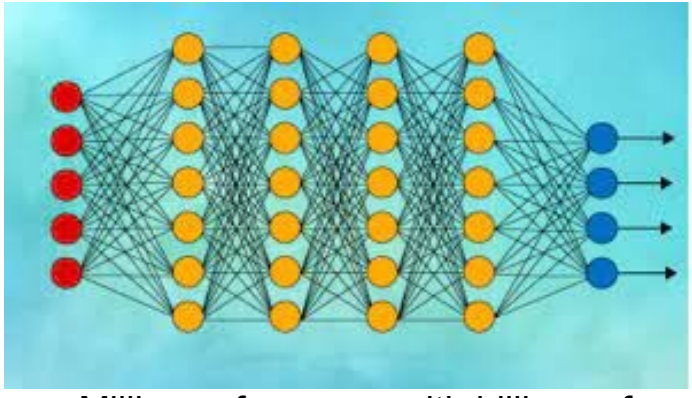
3. **High Energy Consumption:** The energy demands of operating LLMs at full scale are substantial

LLMs and power – Bigger not necessarily better



Weights, biases and intermediate values moved between memory and processing

x



Millions of neurons with billions of connections

=

Billions of bits of data being calculated – moved between memory and processing

More parameters => more data => more bits moved => more energy consumed

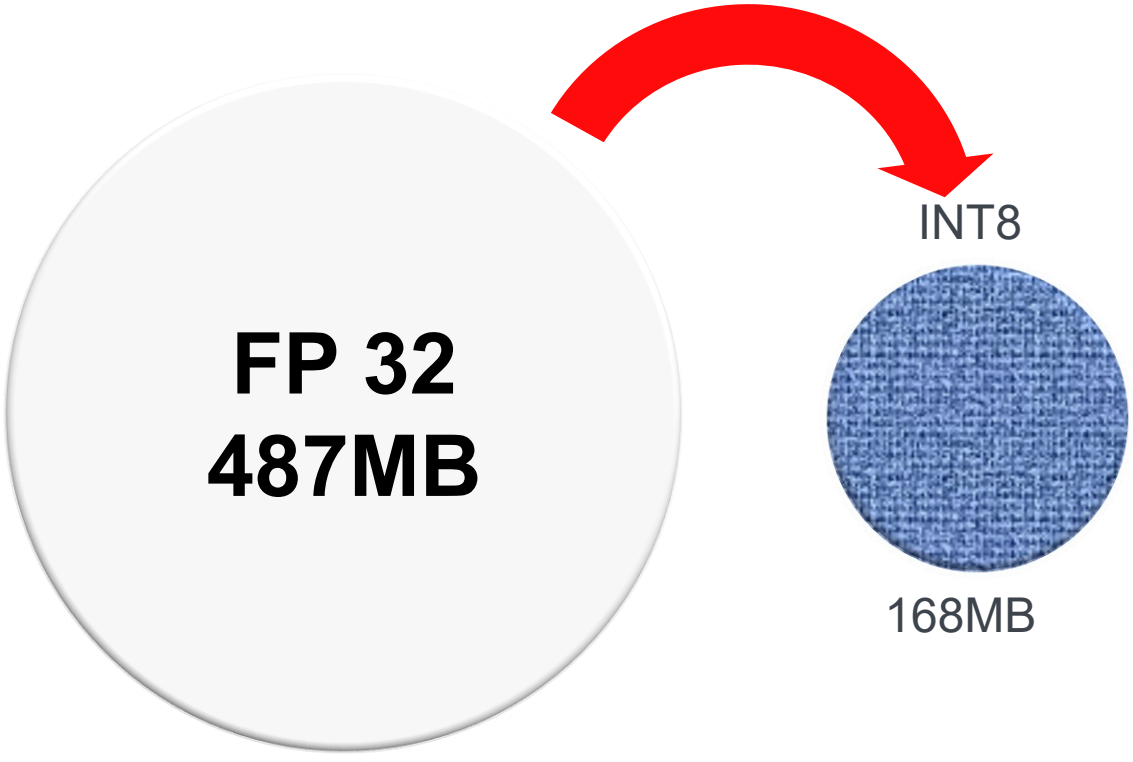
Quantized LLMs for Accessibility and Real-time Processing

1. **Accessibility and Deployment:** Quantization allows LLMs to be more accessible by enabling their deployment on a wider range of devices, including those with limited hardware specifications.

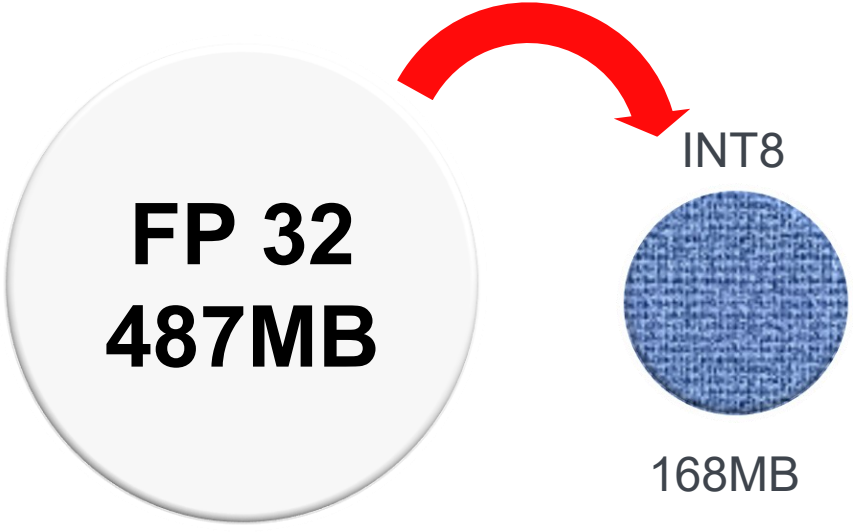
2. **Energy Efficiency:** Reduced resource requirements through quantization lead to lower energy consumption, making LLMs more sustainable and cost-effective for widespread use.

3. **Real-time Applications**: Quantizing LLMs facilitates real-time processing capabilities on edge devices, crucial for applications requiring immediate responses.

Initial ideas



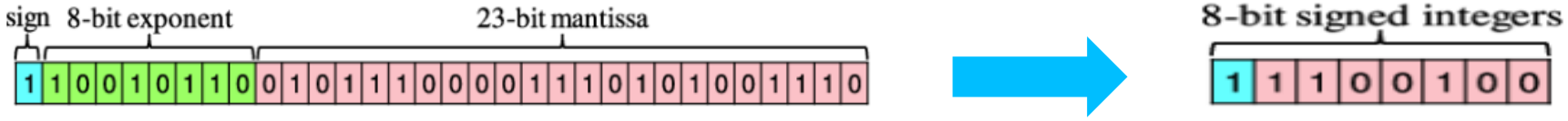
- **Motivation**
- **Basis**
- **Conception Design**
- **Implementation**
- **Evaluation and Verification**
- **Summary and Outlook**



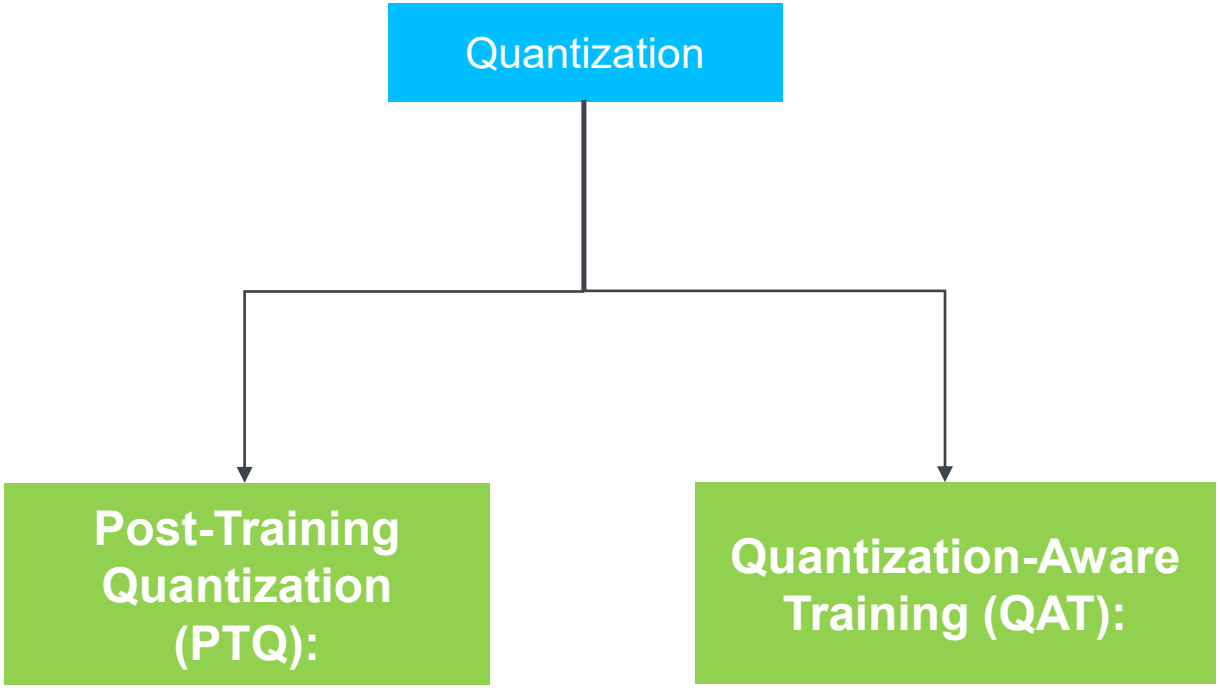
Basis

Quantization

Quantization is an optimization technique that reduces the precision of the numbers used to represent a model's parameters, which are by default 32-bit floating point numbers. The benefits of this optimization are a smaller model size, better portability, and faster computation.



Basis



This method involves converting the weights of a fully trained model to a lower precision format.

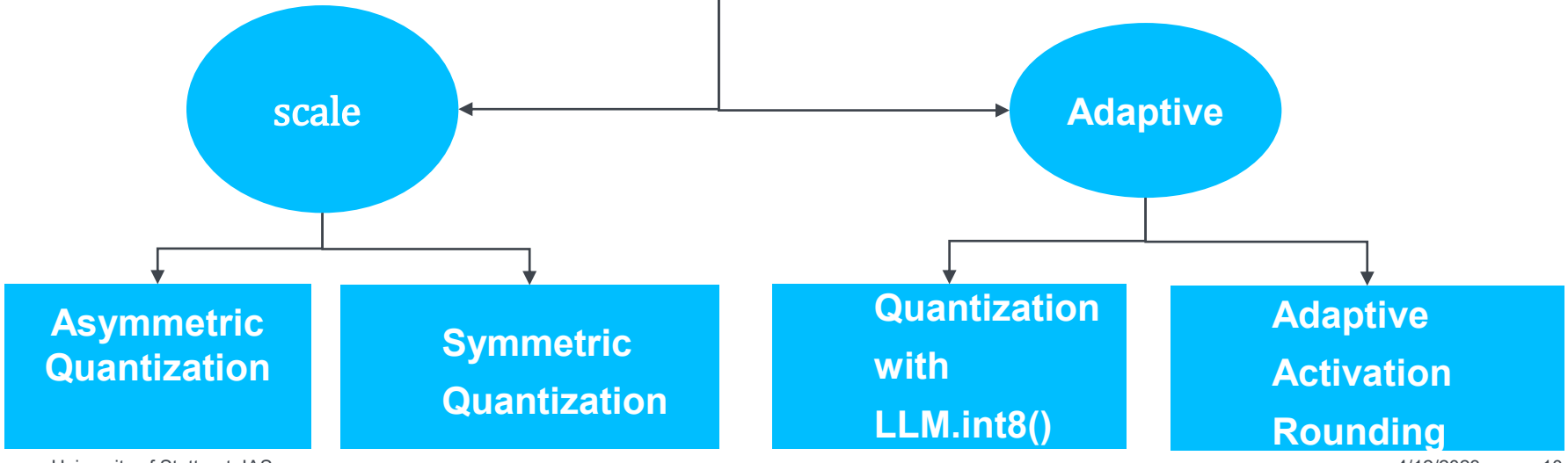
QAT integrates the quantization process into the training phase, either during pre-training or fine-tuning

Basis

Post-Training Quantization (PTQ):

This method involves converting the weights of a fully trained model to a lower precision format.

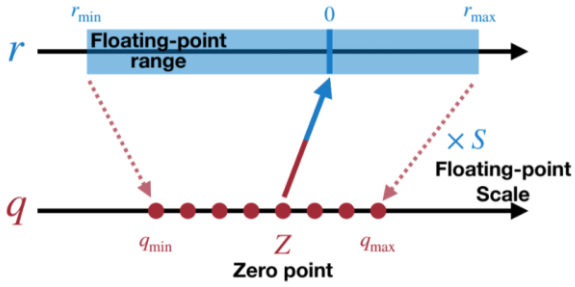
Pros	Cons
simple to implement	drawback potential for performance degradation due to the reduced precision
not require retraining of the model	



Basis

scale

Asymmetric Quantization



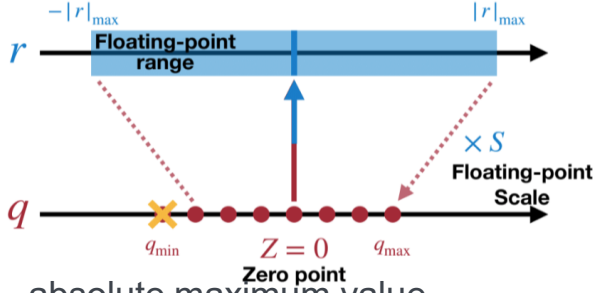
based on the actual minimum (r_min) and maximum (r_max) range of the floating-point numbers

←Scale (S) →

quantized value corresponding to the real value 0. It is determined from the data range and is not necessarily at q_min

←Zero Point (Z)→

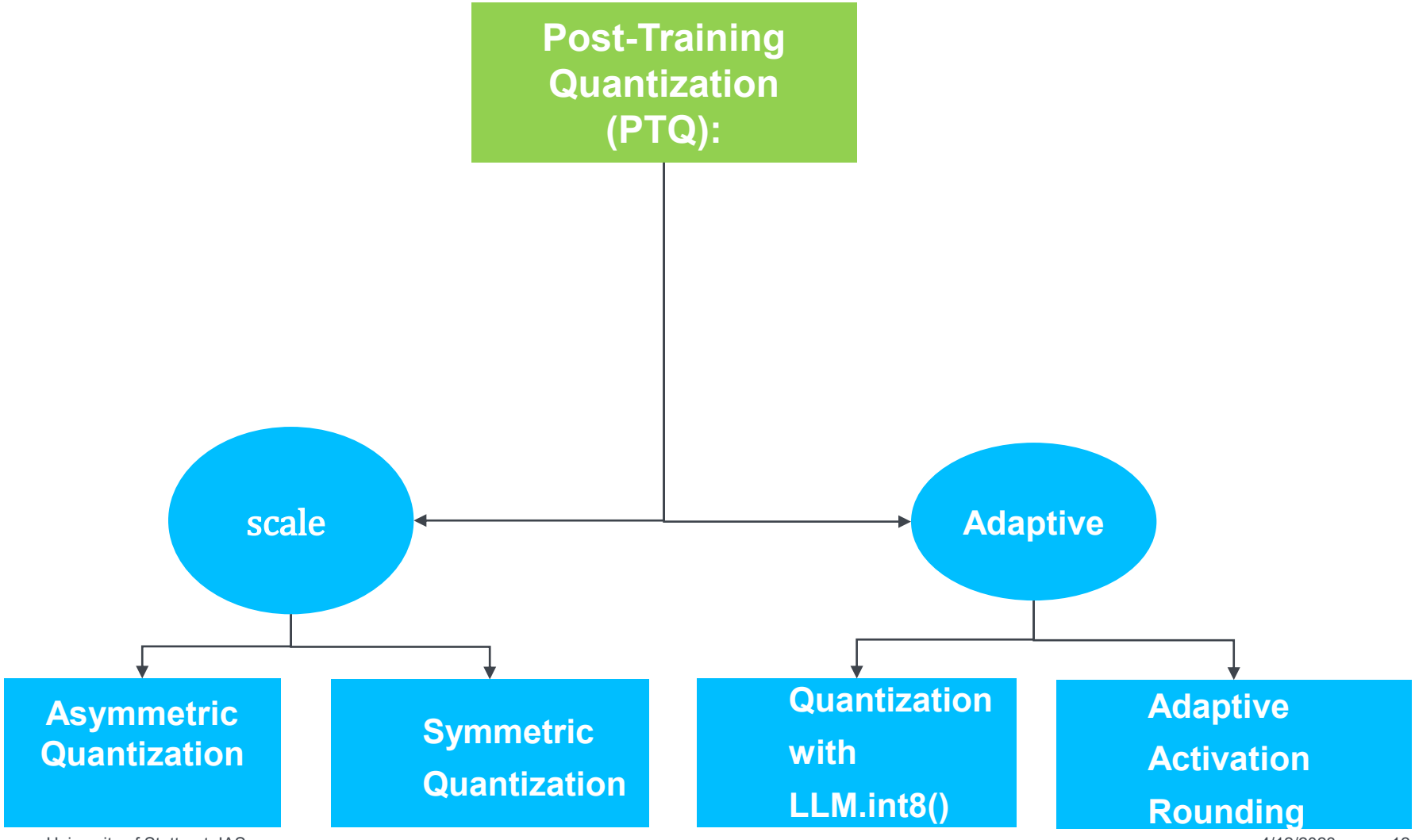
Symmetric Quantization

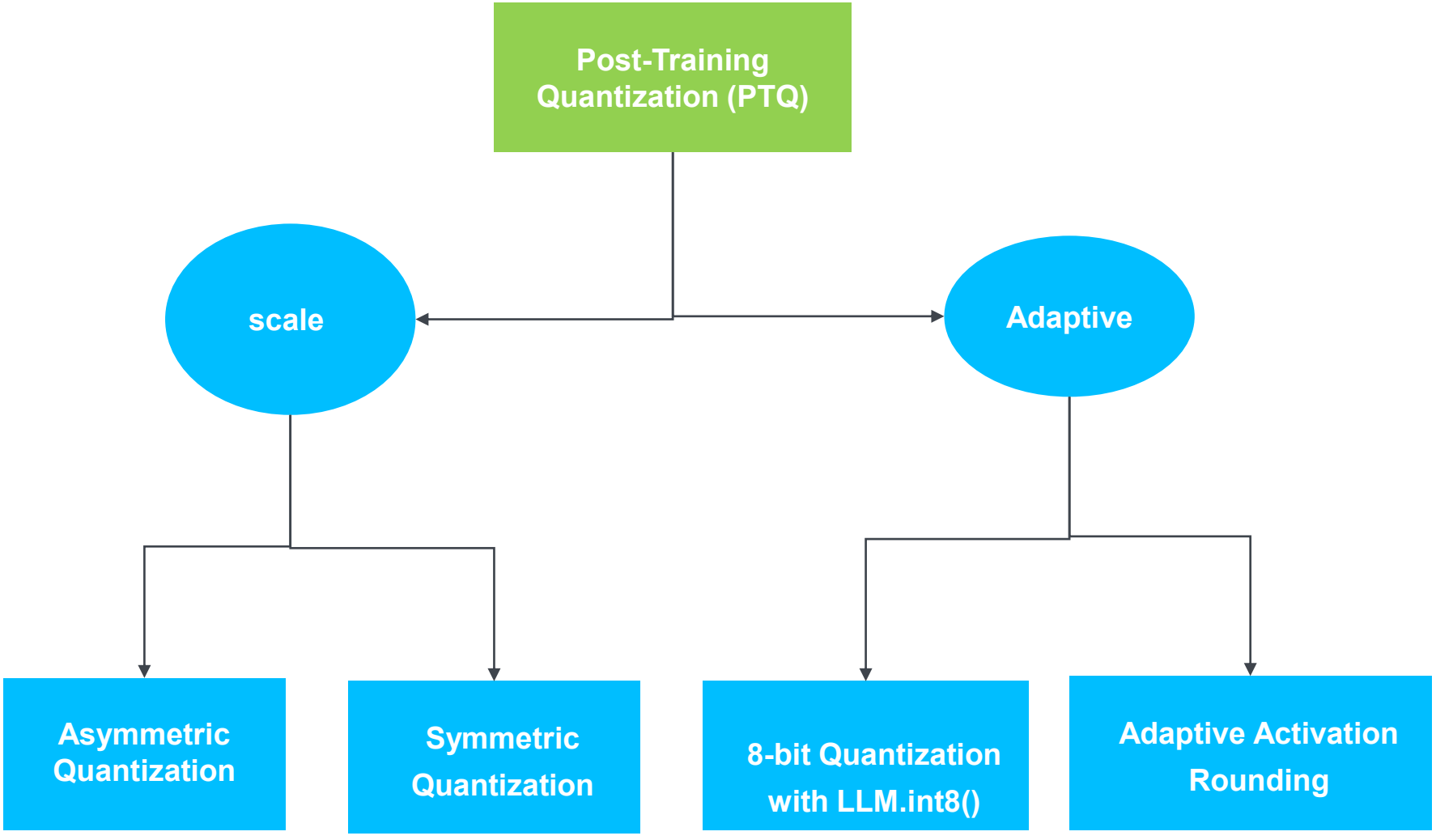


absolute maximum value (|r_max|), creating a symmetric range around zero.

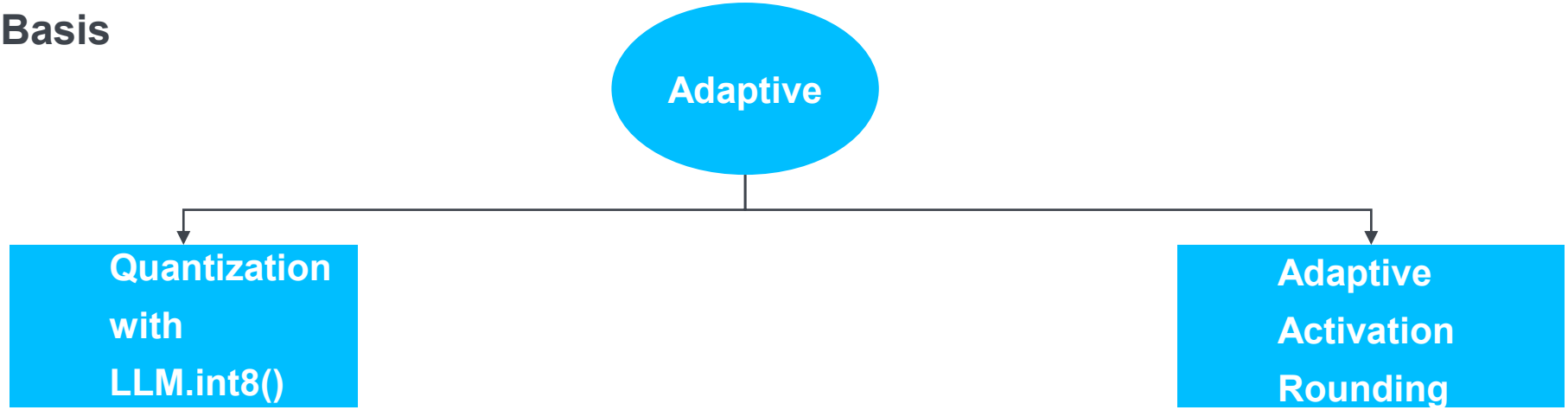
Always at zero in the quantized range, meaning the floating-point zero maps directly to the quantized zero.

Basis


Post-Training Quantization (PTQ):
scale
Adaptive
Asymmetric Quantization
Symmetric Quantization
Quantization with LLM.int8()
Adaptive Activation Rounding


Post-Training Quantization (PTQ)
scale
Adaptive
Asymmetric Quantization
Symmetric Quantization
8-bit Quantization with LLM.int8()
Adaptive Activation Rounding

Basis

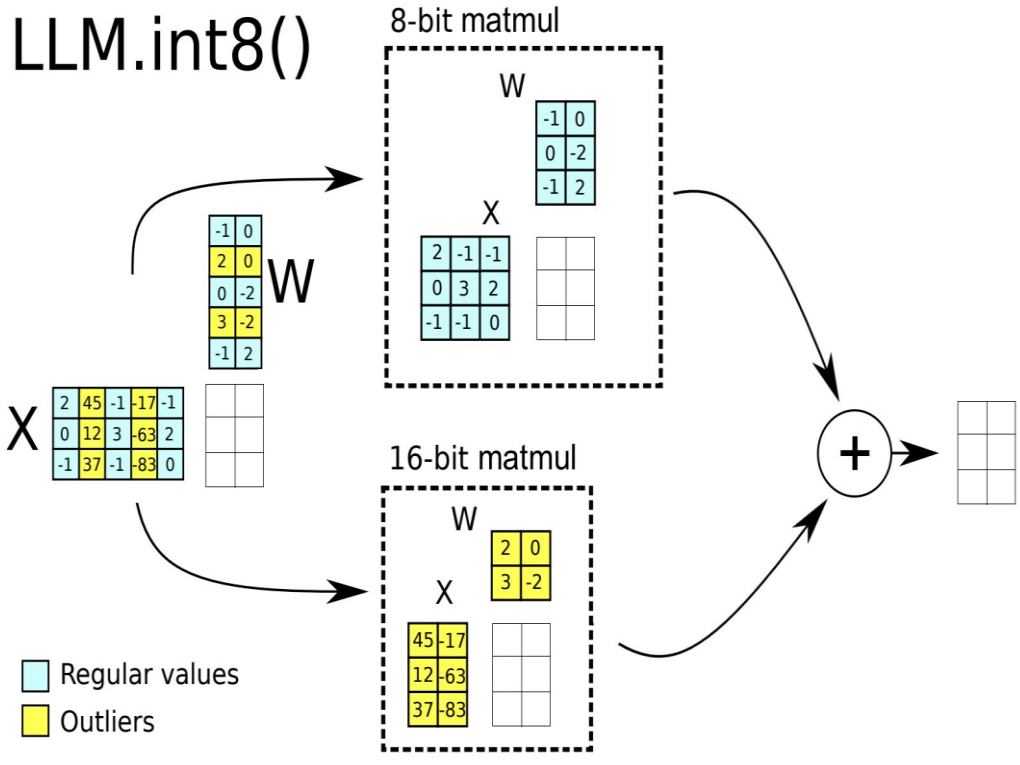

Adaptive
Quantization with LLM.int8()
Adaptive Activation Rounding

8-bit Quantization with LLM.int8()

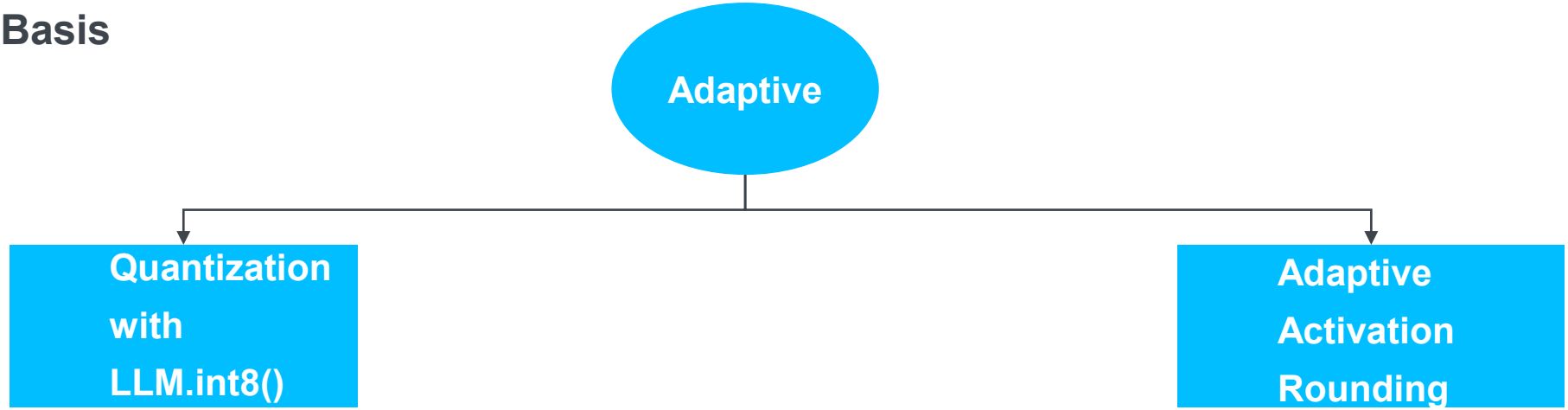
Optimizing with Precision:
Handling Outliers in Quantization
Outlier features

- Exceptional features retain their detail through 16-bit floating-point precision (FP16).

- Transition to 8-bit integer (INT8) format streamlines memory usage while upholdin essential precision.

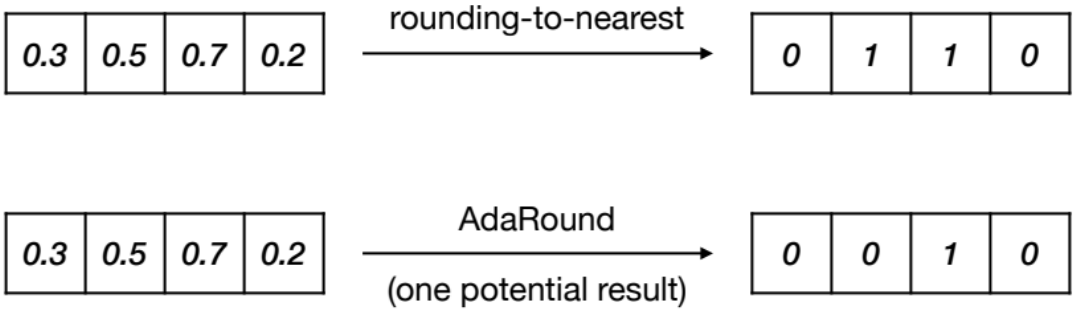


Basis


Adaptive
Quantization with LLM.int8()
Adaptive Activation Rounding

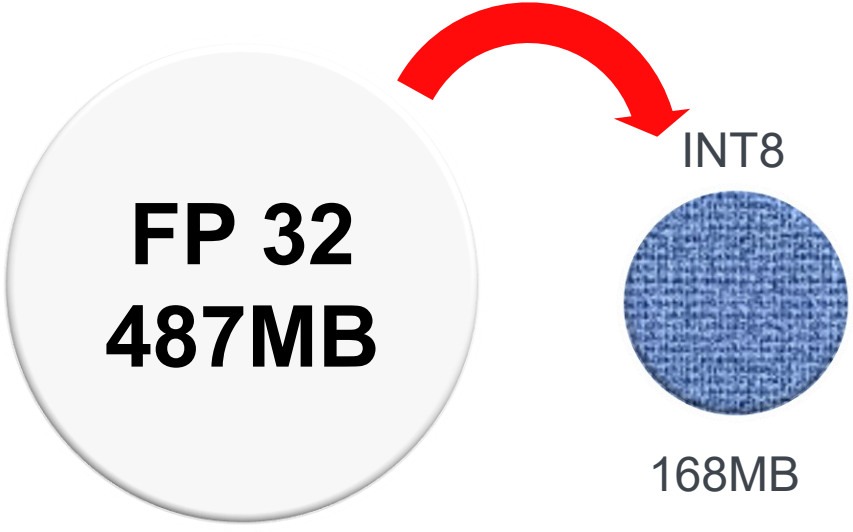
Basis

Adaptive Activation Rounding for Post-Training Quantization

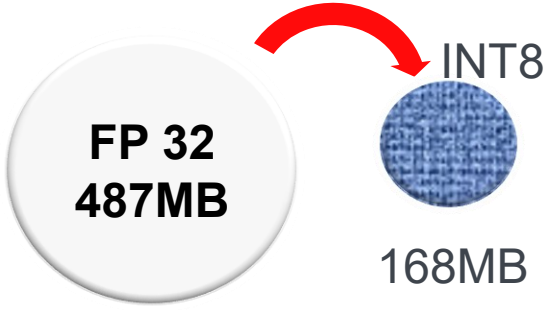


- Simplifies with standard rounding-to-nearest value during post-training quantization.
- Aims to minimize the Frobenius norm, assessing matrix dimensions and calculating error precision.
- Adaptive rounding has been empirically proven to significantly boost model performance and efficiency, particularly in resource-constrained environments such as edge computing and mobil platforms.

- Motivation
- Basis
- Implementation
- Evaluation and Verification
- Summary and Outlook



Implementation



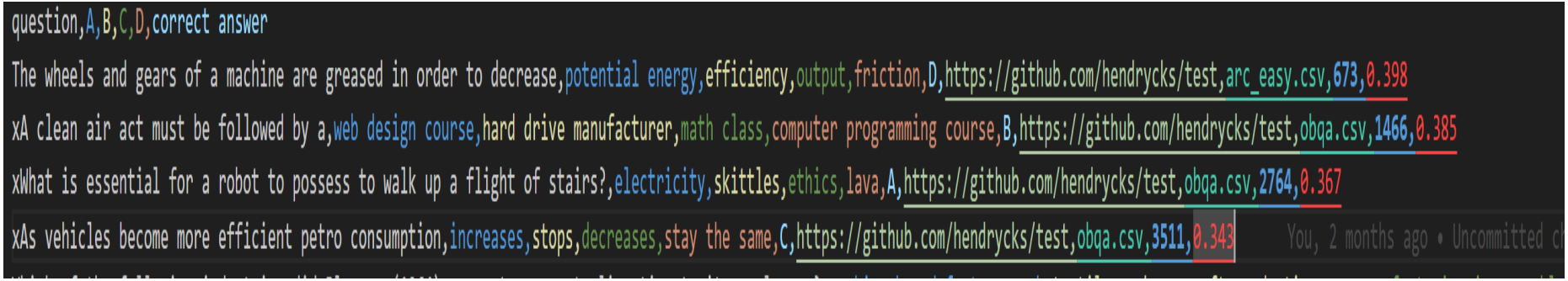
- To quantize pre-trained models there are different libraries and GitHub repositories that explain how to quantize a model in a specific way.
- Also, there are many quantized models available online which can be used.

```
[6] from transformers import AutoModelForCausalLM, AutoTokenizer
    from google.colab import drive
    import torch

    # Mount Google Drive
    device = torch.device('cuda' if torch.cuda.is_available() else 'cpu')
    # Specify your Google Drive path
    model_int8 = AutoModelForCausalLM.from_pretrained(model_id,
                                                      device_map='auto',
                                                      load_in_8bit=True,
                                                      )
    print(f"Model size: {model_int8.get_memory_footprint():,} bytes")
```

```
The `load_in_4bit` and `load_in_8bit` arguments are deprecated and will be removed in the future versions. Please, pass a `BitsAndByt
Loading checkpoint shards: 100% 2/2 [00:04<00:00,  2.13s/it]
Model size: 7,135,051,776 bytes
```

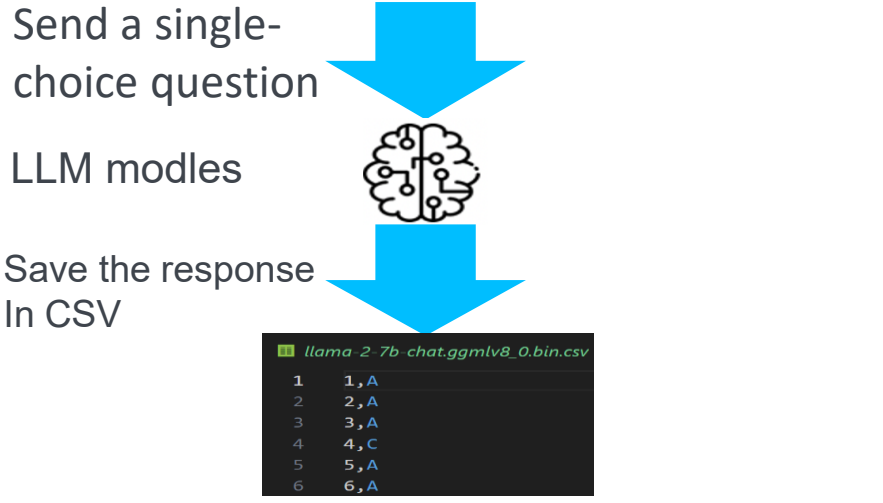
Implementation (Dataset)



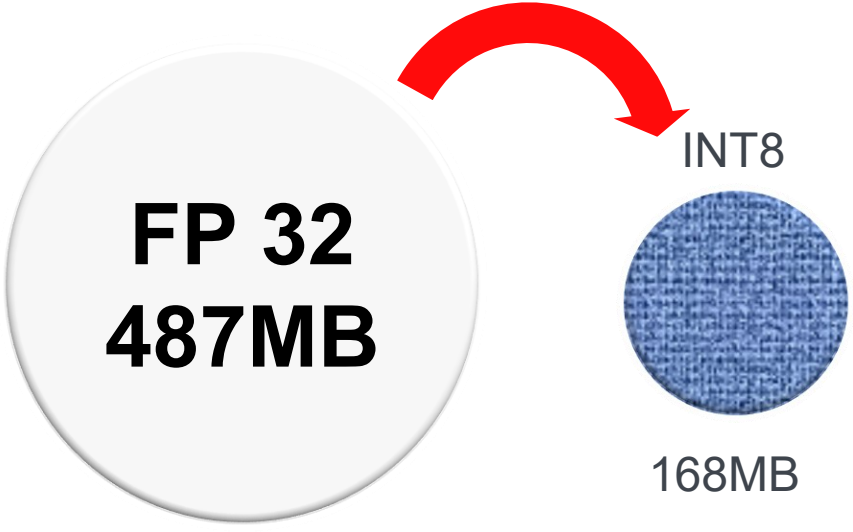
- Objective: To assess the effectiveness of quantized language models in industrial automation applications.
- Dataset Foundation: Utilized a curated set of 875 single-choice questions, extracted from a broader dataset aimed at "MEASURING MASSIVE MULTITASK LANGUAGE UNDERSTANDING.“
- Selection Criteria: Questions were carefully chosen to specifically evaluate the models' comprehension and problem-solving abilities within the industrial automation domain.
- Purpose: This specialized dataset provides a critical benchmarking tool for measuring the nuanced understanding and reasoning capabilities of quantized models in a targeted industrial context.

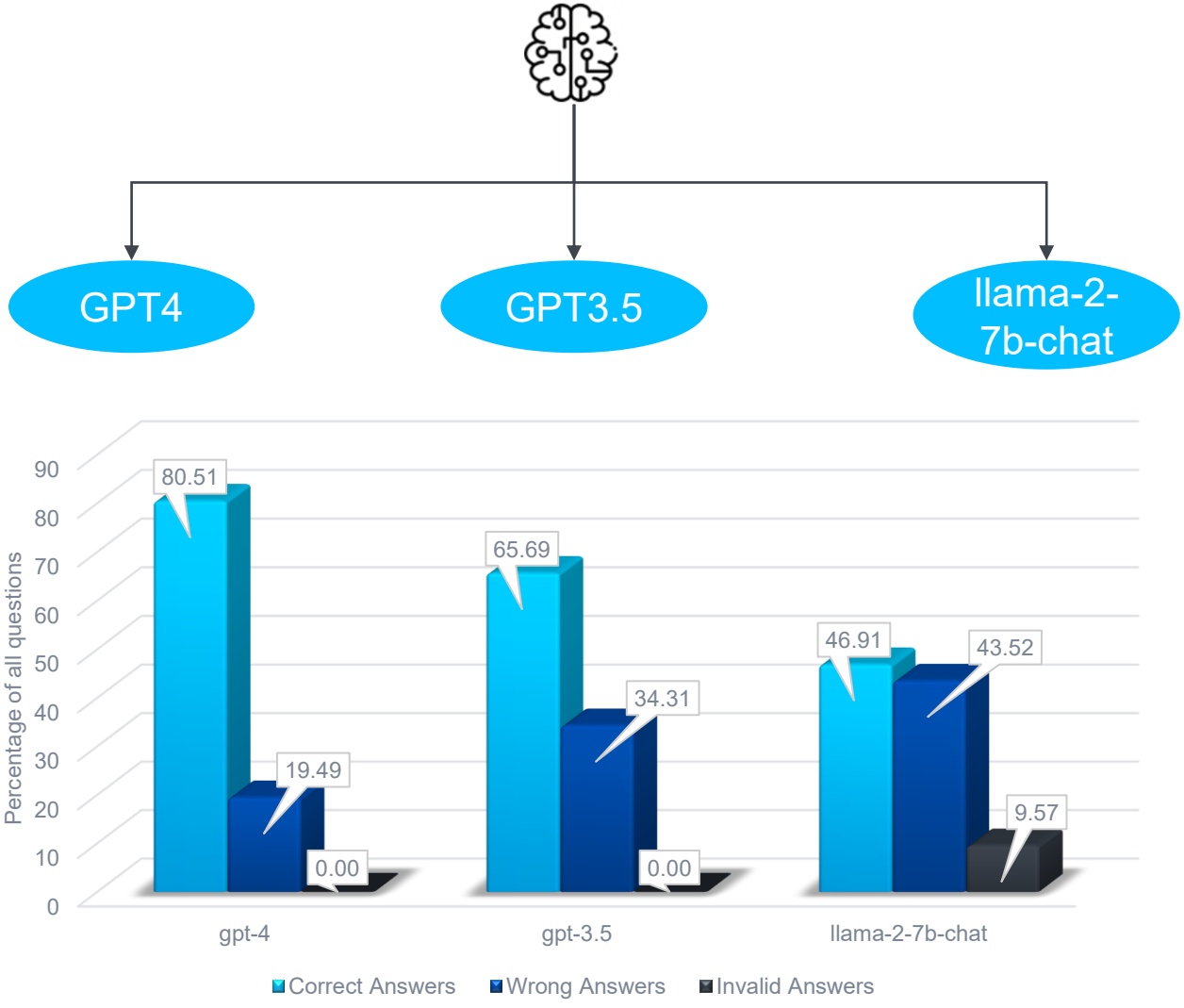
Implementation

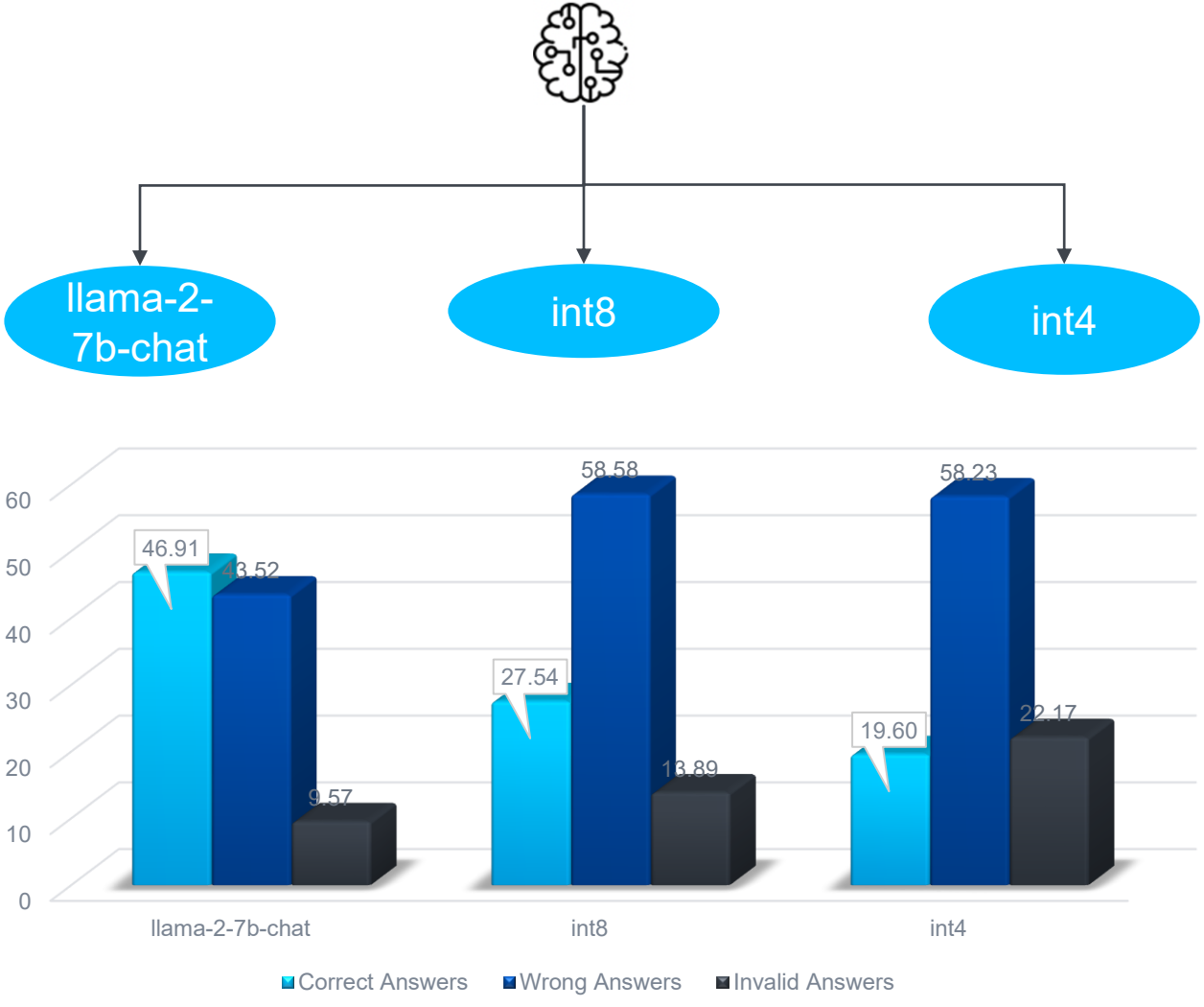
```
question,A,B,C,D,correct answer
The wheels and gears of a machine are greased in order to decrease,potential energy,efficiency,output,friction,D,https://github.com/hendrycks/test,arc_easy.csv,673,0.398
xA clean air act must be followed by a,web design course,hard drive manufacturer,math class,computer programming course,B,https://github.com/hendrycks/test,obqa.csv,1466,0.385
xWhat is essential for a robot to possess to walk up a flight of stairs?,electricity,skittles,ethics,lava,A,https://github.com/hendrycks/test,obqa.csv,2764,0.367
xAs vehicles become more efficient petro consumption,increases,stops,decreases,stay the same,C,https://github.com/hendrycks/test,obqa.csv,3511,0.343      You, 2 months ago • Uncommitted ch
```



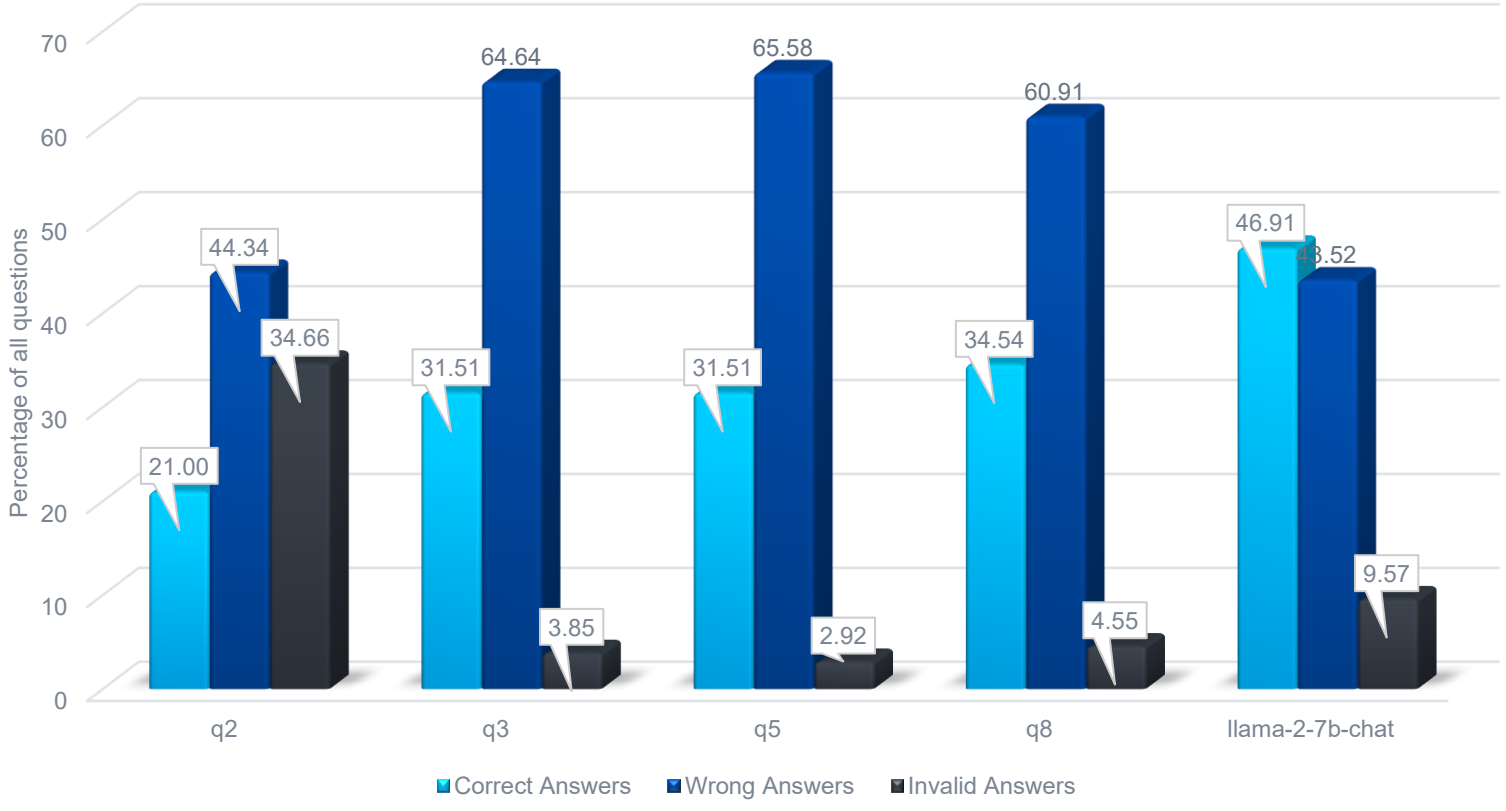
- Motivation
- Basis
- Implementation
- Evaluation and Verification
- Summary and Outlook



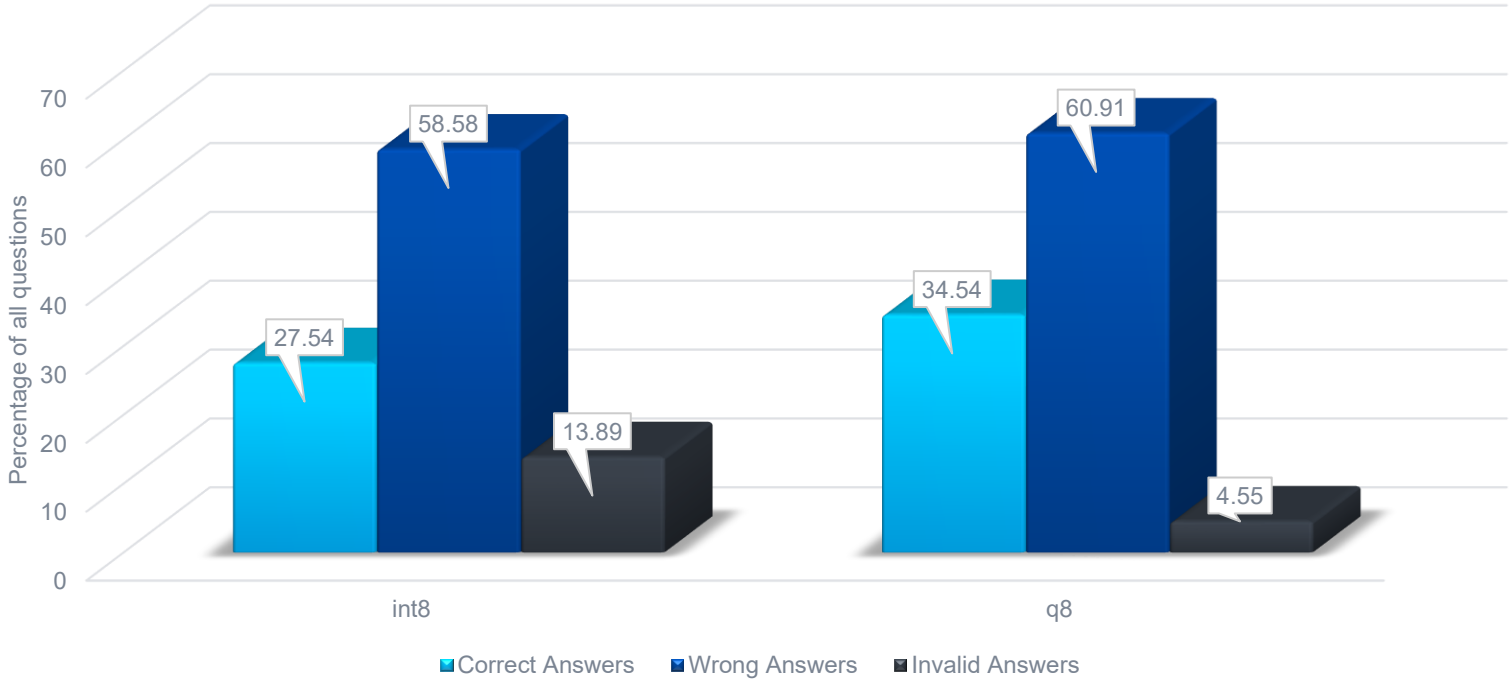

GPT4
GPT3.5
llama-2-7b-chat
Percentage of all questions
90
80
70
60
50
40
30
20
10
0
80.51
19.49
0.00
65.69
34.31
0.00
46.91
43.52
9.57
gpt-4
gpt-3.5
llama-2-7b-chat
Correct Answers
Wrong Answers
Invalid Answers


llama-2-7b-chat
int8
int4
Percentage of all questions
60
50
40
30
20
10
0
46.91
43.52
9.57
27.54
58.58
13.89
19.60
58.23
22.17
llama-2-7b-chat
int8
int4
Correct Answers
Wrong Answers
Invalid Answers

Llama-2-7b-chat and Its Quantized Models (q2k, q3kl, q5kl, q8)



LLM.int8() vs q8



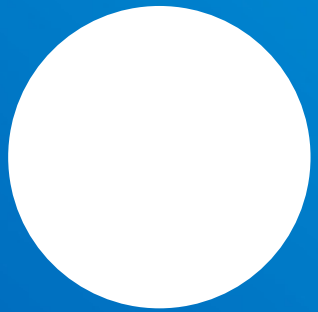
- Ensured fair comparison with consistent parameters such as temperature and prompts.
- Hypothesized q8's potential for higher performance due to alignment with float16 and computational strategy.
- q8 outperforms int8 in accuracy, resulting in fewer invalid responses, confirming
- improved model understanding.

Summary and Outlook

- Investigated quantization on LLMs, emphasizing semantic interpretation for industrial automation.
- Utilized Llama-2-7b-chat model to evaluate efficiency in technical reasoning within industrial contexts.
- Benchmarked against 857 questions from industrial automation papers, measuring accuracy against model size reduction.
- Established that quantization effectively balances model size with performance, maintaining substantial accuracy.
- Confirmed quantized LLMs' viability in industrial automation, interpreting complex data accurately.
- q8 quantization method notably outperformed int8, underscoring the importance of strategic quantization selection.

University of Stuttgart
Institut of Industrial Automation
and Software Engineering

Thank you!



Belal Abulabn

e-mail st18211@stud.uni-stuttgart.de
phone +49 (0) 711 685-
fax +49 (0) 711 685-

University of Stuttgart