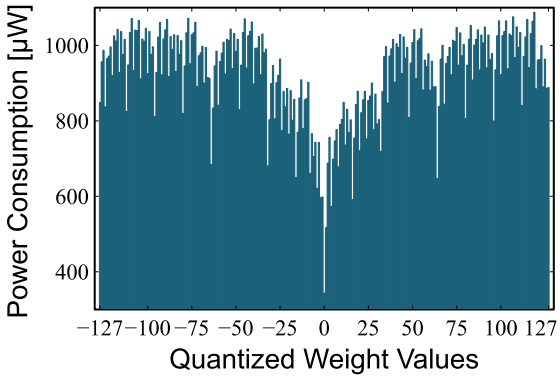
Efficient LLM Inference with Weight Selection

Quantization is one of the techniques for accelerating the execution of large language models (LLMs). However, different quantized values exhibit different computational and power characteristics. A summary of these differences is presented in the table below:

Operation	Fastest Cases	Slower Cases
Add (**x + y**)	x + 0, x + 1	x + (-8)(two's complement overhead)
Mul (**x * y**)	x * 1, x * 2(bit shift)	x * 3, x * 5(full multiplication)
Div (**x / y**)	x / 2, x / 4(bit shift)	x / 3, x / 5(slow)



In this master's thesis, your task is to analyze the impact of different weight values on the latency and power of a decoder-only LLM model. Based on these insights, you will then quantize the model using values that minimize computational latency and power consumption, ultimately achieving a faster and more energy-efficient decoder-only LLM model.

If you are interested in this topic for master thesis, please contact:

Prof. Dr.-Ing. Li Zhang (grace.zhang@tu-darmstadt.de) and Jingcun Wang (jingcun.wang@tu-darmstadt.de) with your CV and transcripts.