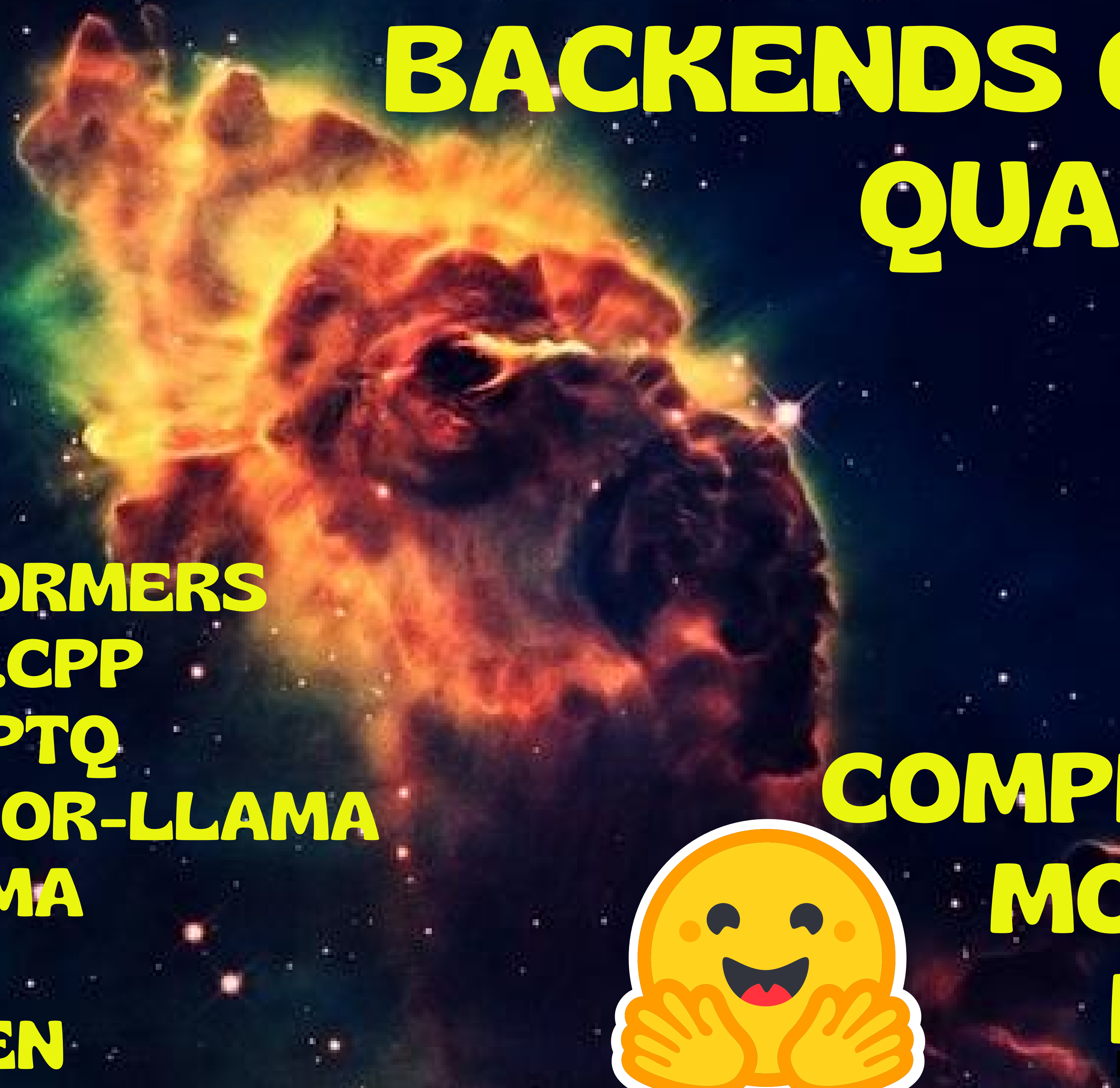
DISCUSSION ON MODEL BACKENDS GPTQ 4-BIT QUANTISATION



OOBABOOGA
TEXT-GEN-UI

- TRANFORMERS
- LLAMA.CPP
- AUTOGPTQ
- GPTQ-FOR-LLAMA
- EXLLAMA
- RWKV,
- FLEXGEN

COMPRESSING THE MODELS AFTER PRETRAINING

HTTPS://GITHUB.COM/INSIGHTBUILDER

CHALLENGE SOLVED: REDUCE SIZE INCREASE SPEED

UNIFIED INTERFACE

HTTPS://GITHUB.COM/OOBABOOGA/ONE-CLICK-INSTALLERS

NOW ITS GPTQ:

- HTTPS://GITHUB.COM/QWOPQWOP200/GPTQ-FOR-LLAMA
- HTTPS://GITHUB.COM/IST-DASLAB/GPTQ
- HTTPS://GITHUB.COM/FPGAMINER/GPTQ-TRITON
- HTTPS://GITHUB.COM/PANQIWEI/AUTOGPTQ

- 4- BIT QUANTIZATION GIVES 95.0% ACCURACY AND 48% OVERALL INFERENCE SPEEDUP,
- 8-BIT QUANTIZED NETWORK GIVES 95.4% ACCURACY AND 39% SPEEDUP.

WHAT IS TRITON?

PROGRAMMING LANGUAGE THAT THE MAIN PREMISE OF THIS PROJECT IS THE FOLLOWING: PRGM BLOCKED INSTEAD OF THREAD PROGRAMMING PARADIGMS BASED ON BLOCKED ALGORITHMS FOR HIGH-PERFORMANCE COMPUTE KERNELS FOR NEURAL NETWORKS.

DISCUSSION

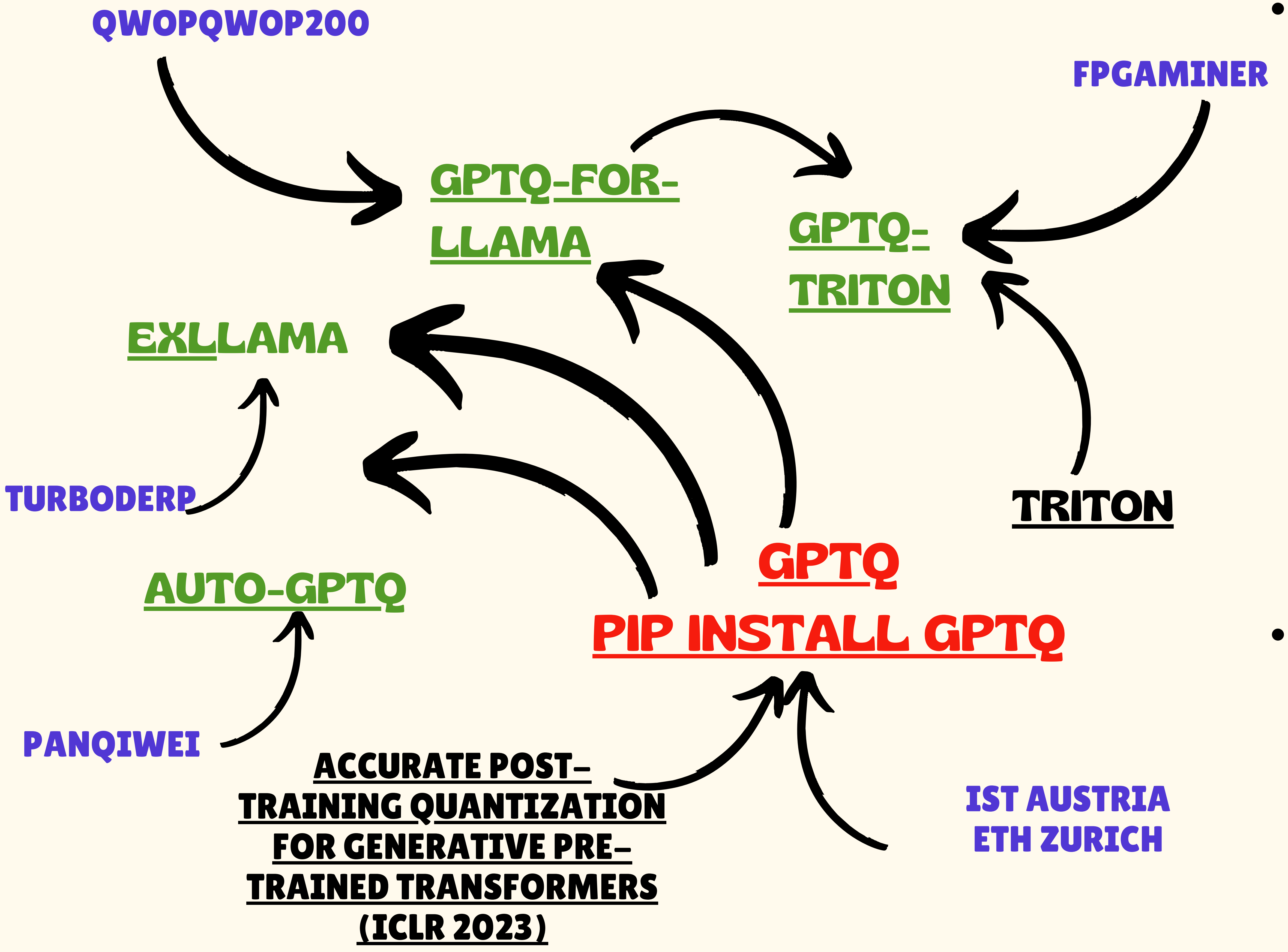
- HTTPS://GITHUB.COM/OOBABOOGA/TEXT-GENERATION-WEBUI/DISCUSSIONS/2740
- HTTPS://WWW.REDDIT.COM/R/LOCALLLAMA/COMMENTS/13UN94P/AUTOGPTQ_VS_GPTQFORLLAMA/

BEFORE:

- HTTPS://GITHUB.COM/TIMDETTMERS/BITSANDBYTES
- HTTPS://GITHUB.COM/HUGGINGFACE/ACCELERATE
- HTTPS://GITHUB.COM/HUGGINGFACE/PEFT

HTTPS://GITHUB.COM/INSIGHTBUILDER

GPTQ EVOLUTION : A MAP



QWOPQWOP200

GPTQ-FOR-LLAMA

GPTQ-TRITON

FPGAMINER

EXLLAMA

TURBODERP

TRITON

GPTQ

AUTO-GPTQ

PIP INSTALL GPTQ

PANQIWEI

ACCURATE POST-TRAINING QUANTIZATION FOR GENERATIVE PRE-TRAINED TRANSFORMERS (ICLR 2023)

IST AUSTRIA
ETH ZURICH

- GPTQ CAN QUANTIZE GPT MODELS WITH 175 BILLION PARAMETERS IN APPROXIMATELY FOUR GPU HOURS, REDUCING THE BITWIDTH DOWN TO 3 OR 4 BITS PER WEIGHT, WITH NEGLIGIBLE ACCURACY DEGRADATION RELATIVE TO THE UNCOMPRESSED BASELINE.
- ALLOWING US FOR THE FIRST TIME TO EXECUTE AN 175 BILLION-PARAMETER MODEL INSIDE A SINGLE GPU FOR GENERATIVE INFERENCE

HTTPS://GITHUB.COM/INSIGHTBUILDER

WHAT BASE GPTQ DOES?

- AN EFFICIENT IMPLEMENTATION OF THE GPTQ ALGORITHM:
- COMPRESSING ALL MODELS FROM THE OPT AND BLOOM FAMILIES TO 2/3/4 BITS, INCLUDING WEIGHT GROUPING:
- EVALUATING THE PERPLEXITY OF QUANTIZED MODELS ON SEVERAL LANGUAGE GENERATION TASKS:
- EVALUATING THE PERFORMANCE OF QUANTIZED MODELS ON SEVERAL ZEROSHOT TASKS:
- A 3-BIT QUANTIZED MATRIX FULL-PRECISION VECTOR PRODUCT CUDA KERNEL:
- BENCHMARKING CODE FOR INDIVIDUAL MATRIX-VECTOR PRODUCTS AND FOR LANGUAGE GENERATION WITH QUANTIZED MODELS:

WHAT DOES ALL THIS MEAN?

MODEL SIZE IS REDUCED

SOME BENCHMARKS

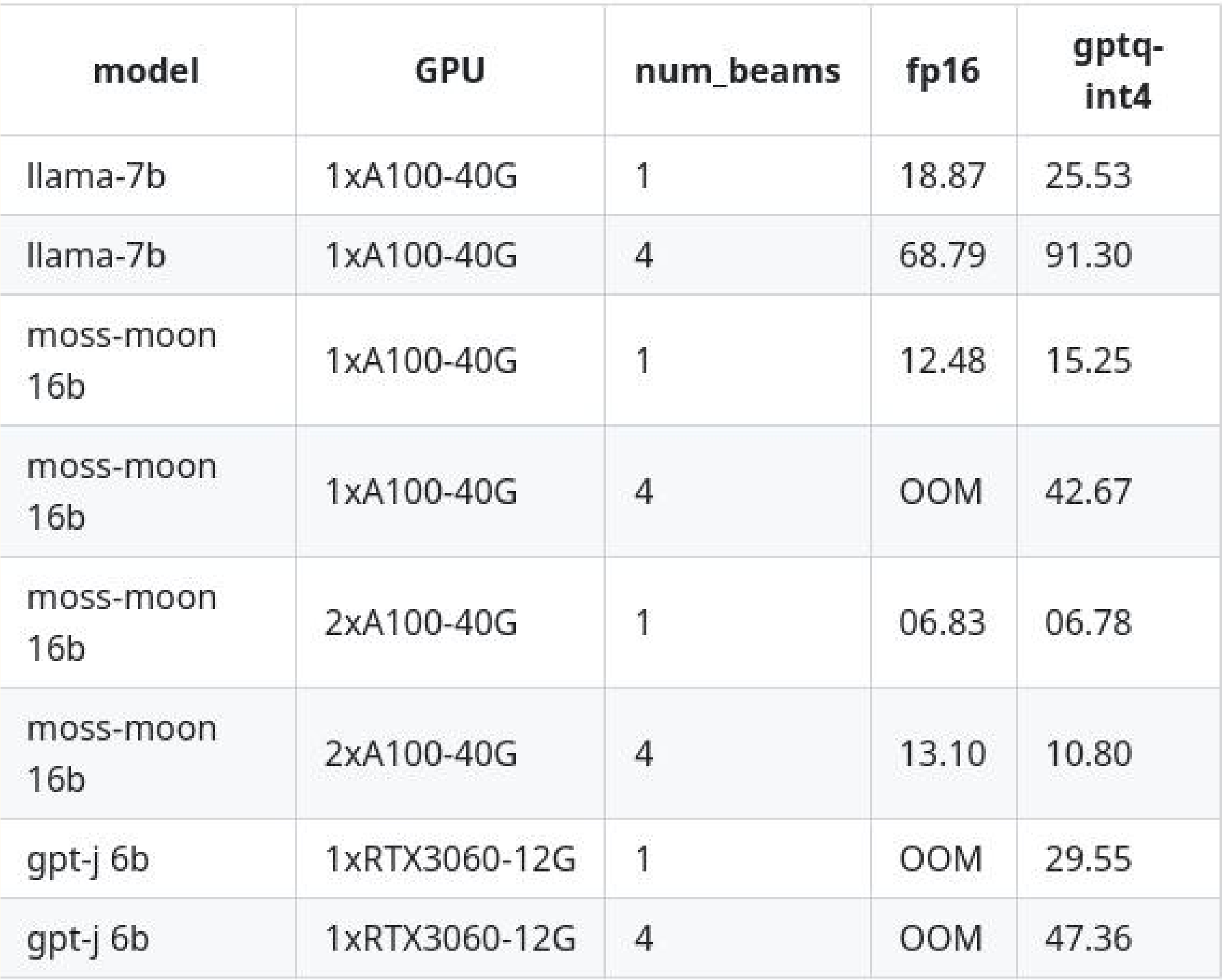
GPTQ TRITON PERFORMANCE

LLaMA-7B	Bits	group-size	memory(MiB)	it/s	Wikitext2	PTB	C4
FP16	16	-	17373	1.64	5.04	7.85	6.99
GPTQ CUDA	4	-1	8805	0.11	5.44	8.24	-
GPTQ Triton	4	-1	6323	1.70	5.44	8.24	7.48

GPTQ FOR LLAMA PERFORMANCE

LLaMA-7B	Bits	group-size	memory(MiB)	Wikitext2	checkpoint size(GB)
FP16	16	-	13940	5.68	12.5
RTN	4	-	-	6.29	-
GPTQ	4	-	4740	6.09	3.5
GPTQ	4	128	4891	5.85	3.6
RTN	3	-	-	25.54	-
GPTQ	3	-	3852	8.07	2.7
GPTQ	3	128	4116	6.61	3.0

AUTO-GPTQ INF SPEEDS



model	GPU	num_beams	fp16	gptq-int4
llama-7b	1xA100-40G	1	18.87	25.53
llama-7b	1xA100-40G	4	68.79	91.30
moss-moon 16b	1xA100-40G	1	12.48	15.25
moss-moon 16b	1xA100-40G	4	OOM	42.67
moss-moon 16b	2xA100-40G	1	06.83	06.78
moss-moon 16b	2xA100-40G	4	13.10	10.80
gpt-j 6b	1xRTX3060-12G	1	OOM	29.55
gpt-j 6b	1xRTX3060-12G	4	OOM	47.36

HTTPS://GITHUB.COM/INSIGHTBUILDER

THANKS FOR WATCHING

REMEMBER TO PRACTICE WITH EXAMPLES

LIKE

SHARE

SUBSCRIBE

HTTPS://GITHUB.COM/INSIGHTBUILDER