

## SmoothQuant and AWQ are widely used:



NVIDIA

FasterTransformer

TRT-LLM

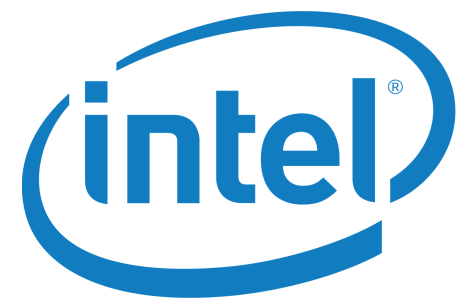
[https://github.com/NVIDIA/FasterTransformer/blob/main/docs/gpt\\_guide.md](https://github.com/NVIDIA/FasterTransformer/blob/main/docs/gpt_guide.md)

<https://github.com/NVIDIA/TensorRT-LLM#key-features>



text-generation-inference

[https://github.com/huggingface/text-generation-inference/tree/main/server/text\\_generation\\_server/utils/awq/quantize](https://github.com/huggingface/text-generation-inference/tree/main/server/text_generation_server/utils/awq/quantize)



Neural Compressor

Q8-Chat

[https://github.com/intel/neural-compressor/blob/master/docs/source/smooth\\_quant.md](https://github.com/intel/neural-compressor/blob/master/docs/source/smooth_quant.md)



Imdeploy

<https://github.com/InternLM/lmdeploy/blob/main/lmdeploy/lite/quantization/awq.py>

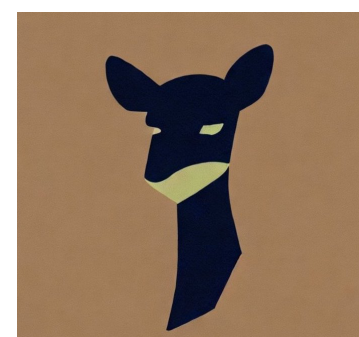


[https://github.com/vllm-project/vllm/blob/main/vllm/model\\_executor/quantization\\_utils/awq.py](https://github.com/vllm-project/vllm/blob/main/vllm/model_executor/quantization_utils/awq.py)



oobabooga/text-generation-webui

<https://github.com/oobabooga/text-generation-webui/blob/main/modules/models.py>



lm-sys/FastChat

<https://github.com/lm-sys/FastChat/blob/main/docs/awq.md>

Replicate

[https://github.com/replicate/vllm-with-loras/blob/main/vllm/model\\_executor/quantization\\_utils/awq.py](https://github.com/replicate/vllm-with-loras/blob/main/vllm/model_executor/quantization_utils/awq.py)