SmoothQuant and AWQ are widely used:

FasterTransformer
TRT-LLM

https://github.com/NVIDIA/FasterTransformer/blob/main/docs/gpt_guide.md

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

Neural Compressor
Q8-Chat

https://github.com/intel/neural-compressor/blob/master/docs/source/smooth_quant.md

vLLM

https://github.com/vllm-project/vllm/blob/main/vllm/model_executor/quantization_utils/awq.py

lm-sys/FastChat

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

text-generation-inference

https://github.com/huggingface/text-generation-inference/tree/main/server/text_generation_server/utils/awq/quantize

lmdeploy

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

oobabooga/text-generation-webui

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

Replicate

https://github.com/replicate/vllm-with-loras/blob/main/vllm/model_executor/quantization_utils/awq.py