Table of Contents

Formats

Safetensors	Safe store for tensors
GGUF	Georgi Gerganov Universal Format (it can mix various precisions)
PT	PyTorch format

Legacy Quantizations (Q4_0, Q4_1, Q5_0, Q5_1, Q8_0, Q8_1): These are simpler, faster methods but may have higher quantization error compared to newer types. K-Quantizations (Q2_K, Q3_K, Q4_K, Q5_K, Q6_K): Introduced in llama.cpp PR #1684, these use super-blocks for smarter bit allocation, reducing quantization error. I-Quantizations (IQ2_XXS, IQ3_S, etc.): State-of-the-art for low-bit widths, using lookup tables for improved accuracy but potentially slower on older hardware.

Opensource models

Stable Diffusion 1

Stable Diffusion XL

https://stability.ai/news/stable-diffusion-sdxl-1-announcement

Stable Diffusion 3

by Stability AI https://stability.ai

Flux.1d

by Black Forest Labs https://bfl.ai

HiDream I1

by HiDream AI https://hidream.org

CLIP (Contrastive Language-Image Pre-training)

Wikipedia

Self-attention Transformer as a text encoder