Quantization LLM — GPTQ, AWQ, EXL2,

16 mai 2026 • Mis à jour le 17 mai 2026 • 15 min de lecture • 3261 mots • 15
vue

Comparez les méthodes de quantization LLM en 2026 : GPTQ, AWQ, EXL2
VRAM requise, la vitesse d'inférence. Guide technique pour choisir la quan

À RETENIR

A retenir -- Quantization LLM 2026

La **quantization LLM** permet de reduire de 50 a 75% la VRAM necessaire pou
la precision des poids de float16 (16 bits) a int4 ou int8. En 2026, quatre form
friendly), **GPTQ** (GPU only, haute qualite), **AWQ** (meilleure preservation de qu
throughput GPU). Pour un deploiement on-premise type ETI, AWQ Q4 ou GG
compromis : 70B model sur un GPU A100 40GB avec moins de 2% de degrad

La **quantization des LLM** est la technique la plus puissante pour democratiser l'ac
sans investir dans des infrastructures GPU prohibitives. Sans quantization, Llama
A100 40GB ou 2x H100 80GB). Avec quantization int4, le meme modele tourne sur

en consumer grade. Cette reduction drastique des besoins en VRAM ouvre des po etaient inatteignables economiquement pour la majorite des organisations. En 202 s'est consolide autour de quatre formats principaux -- GPTQ, AWQ, EXL2 et GGUF distinctes en termes de qualite preservee, de vitesse d'inference, de compatibilite guide technique compare ces quatre approches pour aider les equipes MLOps a c contexte.

Principes de la quantization -- reduire la precision des poids LLM

La ***quantization*** des reseaux de neurones reduit la precision numerique des param la consommation memoire et accelerer les calculs. Pour les LLM, les poids sont ty par valeur) ou bfloat16. La quantization int8 (8 bits, 1 octet) divise la VRAM par 2, e divise par 4.

La challenge de la quantization est de minimiser la degradation de qualite resultan approches naives (arrondir chaque poids au int4 le plus proche) degradent signifi complexes. Les methodes avancees (GPTQ, AWQ) utilisent des algorithmes d'opti precision en ajustant les poids adjacents ou en preservant les poids les plus impor

- **Q8_0 / int8** : degradation quasi-nulle (<0.5% sur les benchmarks standards), re VRAM est la contrainte et que la qualite est prioritaire.
- **Q4_K_M / int4** : degradation faible (1-2% sur les benchmarks standards), reduc qualite/VRAM pour la plupart des deploiements.
- **Q2_K / int2** : degradation significative (5-15%), reduction VRAM par 8. Utile uni (GPU consumer sans suffisamment de VRAM).