AYI NEDJIMI Consultants

# Fine-Tuning LoRA/QLoRA : Guide P

29 April 2026 • Mis à jour le 29 April 2026 • 48 min de lecture

Guide complet fine-tuning LoRA/QLoRA : PEFT, Unsloth, Axolotl, dataset A vLLM/TGI. Comparatif RAG vs fine-tuning.

Le fine-tuning des grands modeles de langage est devenu accessible grace a des reduisent drastiquement les ressources necessaires. LoRA (Low-Rank Adaptation) methodes PEFT (Parameter-Efficient Fine-Tuning) permettent de specialiser des m seul GPU consommateur. Alors qu'un fine-tuning complet de Llama 3 70B necessi dollars de calcul, LoRA permet d'obtenir des resultats comparables en ne modifian VRAM et quelques dizaines de dollars. Ce guide exhaustif couvre les fondements DoRA, la preparation des datasets (formats Alpaca et ShareGPT), les outils d'entra harness), le deploiement (vLLM, TGI), la comparaison avec le RAG et le prompt en adaptateurs. Pour les praticiens de l'IA, les data scientists et les ingenieurs ML, ma fondamentale pour personnaliser les LLM a des cas d'usage specifiques sans exp

À RETENIR

**A retenir :** LoRA ne modifie que 0,1 a 1 % des parametres du modele, reduisan
combine quantification 4 bits et LoRA pour fine-tuner des modeles 70B sur ur
fine-tuning complet pour la plupart des cas d'usage, a une fraction du cout.

**Pourquoi fine-tuner un LLM ?**

Le fine-tuning consiste a poursuivre l'entrainement d'un modele pre-entraine sur
style ou une tache particuliere. Avant d'explorer le comment, il est essentiel de co
tuning est la bonne approche pour votre cas d'usage.

## Quand le fine-tuning est-il necessaire ?

Le fine-tuning est justifie dans plusieurs scenarios. Pour l'adaptation de style ou d
d'une maniere specifique (jargon technique, style juridique, ton de marque) que le
Pour l'apprentissage de formats de sortie complexes, quand le modele doit systen
precis (JSON structure, XML specifique, format de rapport particulier). Pour la spe
maitriser un vocabulaire et des concepts specifiques a un domaine (medical, jurid
que le modele generique fournit. Pour la reduction de latence et de cout, quand vo
les performances d'un modele plus grand sur votre tache specifique. Et pour l'inte
voulez que le modele "connaisse" des informations specifiques a votre organisatio

## Fine-tuning vs RAG vs Prompt Engineering

Le choix entre fine-tuning, RAG et prompt engineering depend du probleme a reso
approche a essayer : il est gratuit (pas de cout d'entrainement), instantane et reve