Résumé

Je recherche un poste d'Ingénieur en Apprentissage Automatique, avec, en particulartié :

- L'entraînement, le peaufinage et le débogage des modèles.
- L'optimisation des modèles pour l'inférence via la quantification, la distillation et l'élagage, ainsi que le développement de noyaux CUDA/Triton.
- Le déploiement et la surveillance des modèles.

En bref : Mon objectif est de construire une infrastructure ML évolutive et efficace !