Summary

I'm seeking a Machine Learning Engineer role, focusing on:

- Training, fine-tuning, and debugging models.
- Optimizing models for inference through quantization and pruning, alongside CUDA/Triton kernel development.
- Deploying and monitoring models.

In short: I'm aiming to construct a scalable and efficient ML infrastructure!