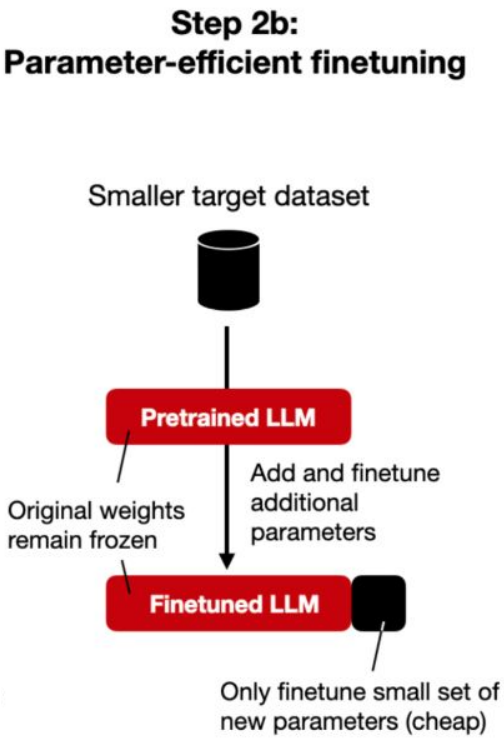
LLM Detoxification with SFT Trainer & TrustyAI Detoxify

Christina Xu

What is SFT Trainer ?

- Pre-training LLMs enables them to predict the next token but can lead to undesirable outputs
- Supervised fine-tuning trains them into useful assistants / chatbots
- SFT Trainer simplifies instruction tuning and supports PEFT including QLoRA

How does QLoRA work ?



- With LoRA, the base model is kept in 32 or 16 bits in memory
- QLoRA compresses the model down to 8 or 4 bits

How does TrustyAI Detoxify work ?

- Rephrases text according to toxicity scores
- Toxicity scores are calculated based on the difference of next token probability distribution between two models
- Mask tokens with the highest toxicities
- Predict non-toxic alternatives for masked tokens