Safe LLM adaptation framework

Kryzhanovskiy Maxim

October 24, 2024

Scope

LLM training stages

LLM usually undergoes following stages:

- Pre-train
- Alignment
  - SFT
  - Preference Optimization

HHH

Safety LLM

Alignment process should make LLM to fit HHH:

- **H**elpful
- **H**onest
- **H**armless

LLM adaptation approaches

To adapt LLM to narrow domain language it should undergo all the same stages that base model did but with specific language. However, lack of labeled data usually allows only to perform pre-training through following approaches:

- Fine-tuning
- Parameter efficient fine-tuning
  - LoRA
  - DoRA
  - ptuningV2

But with fine-tuning HHH-criteria of our model struggles.