Safe LLM regularization

Vorontsov Konstantin, Ischenko Roman, Kryzhanovskiy Maxim

November 22, 2024

Problem formulation

Stages of LLM training

- Pretrain
- Alignment
  - Supervised Fine-tuning
  - Preference Optimization

LLM Adaptation

LLM Adaptation approaches

- Domain specific pretrain
  - Fine-tuning
  - PEFT methods
- Domain specific alignment

Fine-tuning kills Alignment

Fine-tuning of LLM kills alignment

- Keeping LLMs Aligned After Fine-tuning: The Crucial Role of Prompt Templates
- Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!

Red teaming

Red teaming

Red teaming - research field that studies approaches for creating adversarial attacks on LLM to compromise its safety (**red prompts**)

Re teaming datasets

- ALERT
- Thoroughly Engineered Toxicity

Approach for toxicity robustness evaluation

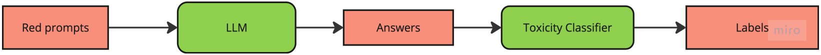


Figure 1: Toxicity robustness evaluation framework

KL divergence

KL Divergence

In general, similarity between probability distributions

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} \, dx \tag{1}$$

Loss with KL divergence regularization

Loss with KL regularization

Designing loss with KL regularization.

- $$\mathcal{L}_{\text{causal}} = -\sum_{t=1}^{T} \log P(x_t | x_{<t}; \theta) + Reg(\theta, \theta*) \tag{2}$$

- $$Reg(\theta, \theta*) = \gamma^t D_{KL}(\theta || \theta*) \tag{3}$$

  $$\theta - current\ \ model\ \ parameters$$

  $$\theta * - base\ \ model\ \ parameters$$

  $$\gamma - decay\ \ rate,\ \ t - epoch$$

Task

General task formulation

To adapt LLM for scientific domain using fine-tuning while keeping it safe with no alignment data given apriori

Data

Data

- Fine-tuning data - Arxiv collection and Elibrary
- Red prompts - ALERT and TET datasets

Experimental setup

Setup

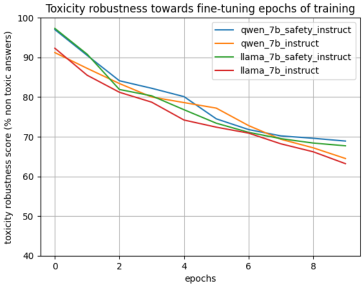
- Hardware - GPU NVIDIA A100
- 10 epochs of training

Evaluation

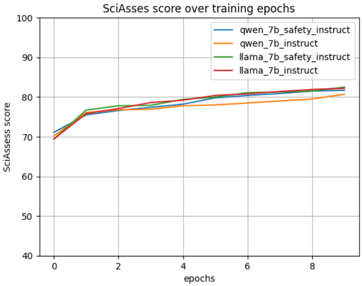
Evaluation

- Domain-specific evaluation - SciAssess
- Toxicity Evalution - as described previously using ALERT and Red Teaming datasets, Llama7b-instruct as toxicity evaluator

Classic fine-tuning



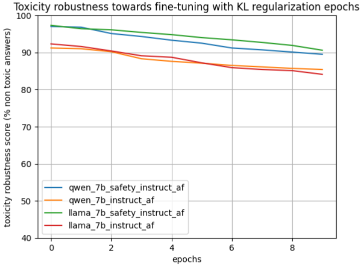
(a) Results during tuning on toxicity robustness



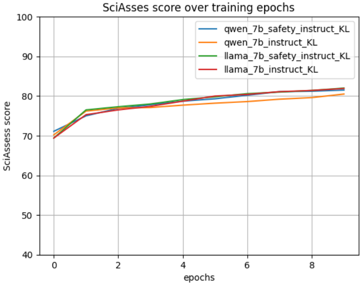
(b) Results during tuning on SciAssess

Figure 2: Classic Tuning

Fine-tuning with Regularization



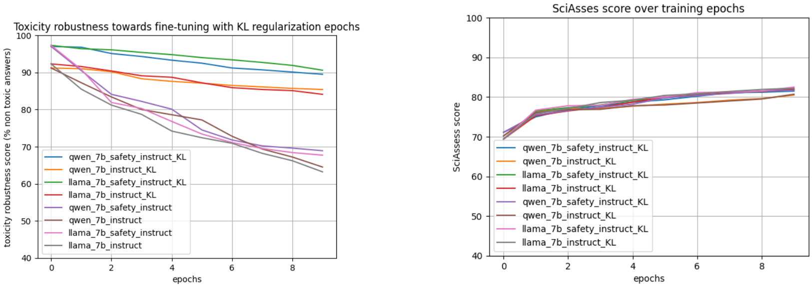
(a) Results during tuning on toxicity robustness



(b) Results during tuning on SciAssess

Figure 3: Tuning with regularization

Comparison



(a) Results during tuning on toxicity robustness

(b) Results during tuning on SciAssess

Figure 4: Comparison

Discussion

Advantages

- More robust towards toxicity
- Comparable results with classic fine-tuning

Disadvantages

- Memory usage
- More resources

Literature

- ALERT: A Comprehensive Benchmark for Assessing Large Language Models' Safety through Red Teaming
- Realistic Evaluation of Toxicity in Large Language Models
- GPT (Generative Pre-trained Transformer) A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions