Safe LLM adaptation through regularization

A Preprint

Maxim Kryzhanovskiy
Calculus Mathematics and Cybernetics
Lomonosov Moscow State University
kryzhanovskiymax@mail.ru

Roman Ischenko
Calculus Mathematics and Cybernetics
Lomonosov Moscow State University
rois@mlsa-iai.ru

Abstract

The adaptation of Large Language Models (LLMs) is a critical yet challenging process, typically comprising two key stages: pretraining and alignment. During the alignment stage, models are refined to produce human-like responses and to ensure safety, particularly by mitigating harmful or toxic outputs. However, when well-aligned models undergo further pretraining in a new domain, they often lose these safety attributes, compromising their alignment. In this study, we explore various strategies to address this issue and propose a novel regularization method designed to preserve alignment during the pretraining phase. Our method significantly mitigates the degradation of alignment, ensuring that models retain their safety and alignment qualities even after domain-specific pretraining.

Keywords Large Language Model · Alignment · Pretrain · Regulartization

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities across diverse applications, from answering complex queries to generating creative content. The adaptation process for these models generally consists of two key stages: pretraining and alignment. Pretraining involves exposing the model to a vast corpus of data to develop a foundational understanding of language, while alignment fine-tunes the model to produce safe, human-aligned outputs by addressing issues like toxicity, bias, and coherence in responses.

Despite the progress achieved through alignment techniques, an emerging challenge arises when aligned models undergo further pretraining for domain-specific tasks or new knowledge integration. This process often compromises the alignment gains, leading to unintended consequences such as reintroducing toxic behaviors, losing human-like response quality, or exhibiting biases. The degradation of alignment poses a significant problem, especially for real-world deployments where safety and reliability are paramount.

Addressing this issue is crucial for ensuring that LLMs maintain their safety and usability across diverse domains without compromising their ability to learn new knowledge effectively. In this work, we investigate the root causes of alignment degradation during pretraining and explore strategies to mitigate its impact. To this end, we propose a novel regularization method specifically designed to preserve the alignment of LLMs even when retrained for new domains.

Our contributions are as follows:

- We identify and analyze the factors leading to alignment degradation during pretraining.
- We propose a regularization technique that significantly mitigates this effect, ensuring alignment consistency.
- We demonstrate the efficacy of our method through empirical evaluations across multiple domains and scenarios.

By addressing this critical issue, our work bridges a gap in the adaptation process of LLMs, paving the way for safer and more reliable language models that can seamlessly integrate domain-specific knowledge without losing their alignment attributes.

2 Related Works

The development and adaptation of Large Language Models (LLMs) have been the focus of extensive research in recent years, spanning several interconnected domains. Below, we briefly discuss the relevant advancements in LLM architectures, safety evaluation through red teaming, alignment strategies, and pretraining methodologies.

LLM Architectures and Main Models. The evolution of LLMs has been driven by innovations in transformer-based architectures, starting with models like GPT-2 and BERT and culminating in cutting-edge systems such as GPT-4, PaLM, and LLaMA. These models leverage self-attention mechanisms to capture complex dependencies in text, enabling state-of-the-art performance across a wide range of tasks. Key innovations, such as sparse attention, multi-modal integration, and scaling laws, have pushed the boundaries of what LLMs can achieve. The role of massive pretraining datasets and optimized training pipelines has been critical in achieving high levels of fluency, comprehension, and contextual understanding.

Red Teaming: Breaking LLM Safety. Red teaming is an essential methodology for evaluating and improving the safety of LLMs. By intentionally probing models with adversarial inputs, researchers aim to uncover vulnerabilities, such as generating toxic, biased, or harmful content. This approach has been instrumental in identifying the limitations of current alignment techniques and in driving the development of more robust safety mechanisms. Recent work in this area has highlighted the persistent challenge of maintaining safety under distributional shifts, such as those encountered during domain-specific pretraining.

Alignment: Techniques and Challenges. Alignment focuses on fine-tuning LLMs to produce responses that are safe, coherent, and aligned with human values. Methods like Reinforcement Learning from Human Feedback (RLHF) have become standard for aligning models with human preferences. Other approaches involve dataset curation, rule-based postprocessing, and embedding safety constraints directly into model architectures. Despite these advancements, maintaining alignment during subsequent training phases, such as domain-specific pretraining, remains a significant challenge. Misaligned models can produce unintended outputs that undermine user trust and the model's overall utility.

LLM Pretraining. The pretraining stage forms the backbone of LLM development, wherein models are exposed to extensive corpora to develop a broad understanding of language. Techniques such as masked language modeling and autoregressive modeling have been employed to pretrain models effectively. However, this stage is not without challenges, as the data used may inadvertently encode biases, and the scale of training often results in the emergence of unpredictable behaviors. Moreover, pretraining aligned models for new domains often disrupts their safety mechanisms, creating a conflict between acquiring domain-specific knowledge and retaining alignment.

3 Safe Regularization

3.1 Toxicity robustness evaluation

For testing our approach and discovering regularization effect on toxicity robustness we need framework that provides with information about LLM robustness.

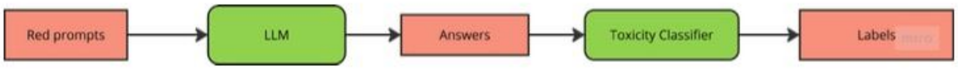


Рис. 1: Toxicity robustness evaluation framework

We collect dataset that consists from red prompts from two public datasets: ALERT and TeT. This prompts are passed to the LLM and obtained answers are labeled with binary toxicity classifier.

3.2 LLM regularization

The Kullback-Leibler (KL) divergence is a measure of how one probability distribution differs from a reference distribution. It is widely used in machine learning to quantify the distance between two distributions. The KL divergence between two distributions P and Q is defined as:

$$D_{KL}(P \parallel Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}, \tag{1}$$

where $P(x)$ and $Q(x)$ are the probabilities assigned to event x by the respective distributions. In the context of LLM training, P often represents the target distribution (e.g., human-aligned outputs), while Q represents the model's predicted distribution. KL divergence is particularly useful for regularization, as it encourages the model's predictions to stay close to a desired reference distribution.

During pretraining, LLMs are optimized to predict the next token in a sequence based on preceding tokens. This is achieved by minimizing the negative log-likelihood (NLL) of the target tokens given the input context. The pretraining loss is formulated as:

$$\mathcal{L}_{pretrain} = -\sum_{i=1}^{N} \log P_\theta(x_i \mid x_{<i}), \tag{2}$$

where x_i is the i-th token in the sequence, $x_{<i}$ represents all tokens preceding x_i, P_θ is the probability distribution over the vocabulary predicted by the model with parameters θ, and N is the sequence length.

This loss function encourages the model to assign higher probabilities to the correct next tokens, thereby learning a robust language representation. However, without additional constraints, this optimization can lead to undesirable outcomes, such as the degradation of alignment properties when fine-tuning or retraining in a new domain.

In our work we provide following approach for regularization.

$$\mathcal{L}_{pretrain} = -\sum_{i=1}^{N} \log P_\theta(x_i \mid x_{<i}) + \gamma^t D_{KL}(\theta \parallel \theta_{old}), \tag{3}$$

γ - regularization coefficient, t - epoch number, θ_{old} - parameters of the model on previous iteration.

4 Experimental setup

To evaluate the effectiveness of our proposed regularization method, we conducted experiments using two state-of-the-art Large Language Models (LLMs): LLaMA 7B and Qwen 7B. These models were selected due to their widespread adoption and strong performance across various natural language processing tasks.

For the evaluation of domain-specific capabilities, we utilized the SciAssess benchmark, which is specifically designed to test a model's proficiency in scientific and technical domains. SciAssess includes tasks such as domain-specific text generation, question answering, and reasoning, providing a comprehensive assessment of the models' ability to adapt to specialized fields while maintaining alignment properties.

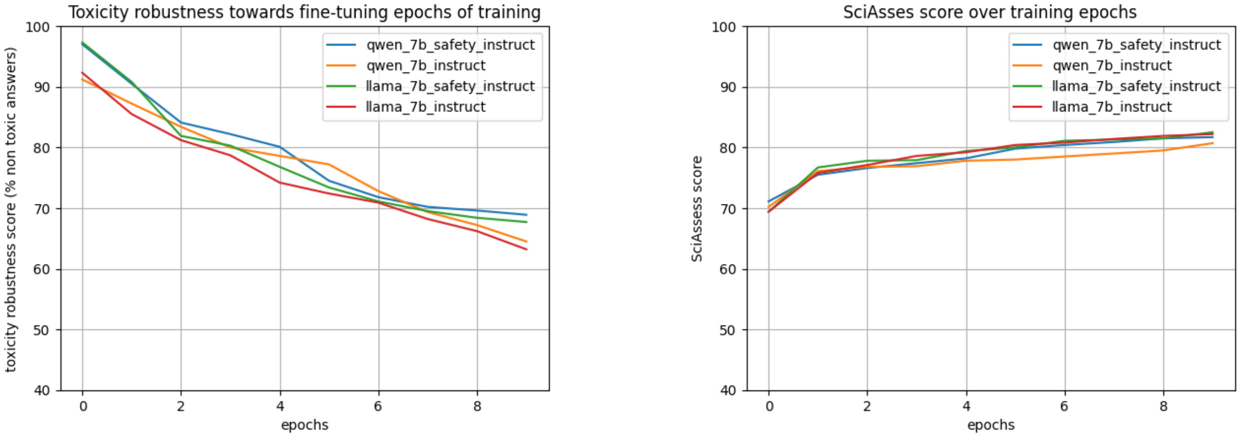
Our experiments involved the following key steps:

1. Baseline Evaluation: Both LLaMA 7B and Qwen 7B were evaluated on SciAssess without applying any domain-specific pretraining or regularization. This established a performance baseline for comparison.
2. Domain-Specific Pretraining: The models were further pretrained using scientific and technical corpora to enhance their domain knowledge.
3. Regularized Training: Our proposed regularization method was applied during domain-specific pretraining to mitigate alignment degradation.
4. Post-Training Evaluation: The models were re-evaluated on SciAssess to measure the impact of domain-specific pretraining with and without regularization. Additionally, alignment properties were assessed to ensure that safety and human-like response qualities were preserved.
5. Toxicity robustness evaluation. On each epoch of training evaluating LLM capabilities of withstanding red prompts.

5 Experiments

5.1 Classic pretrain approach

In this section we provide results of classic tuning.



(a) Toxicity robustness evalution through epochs

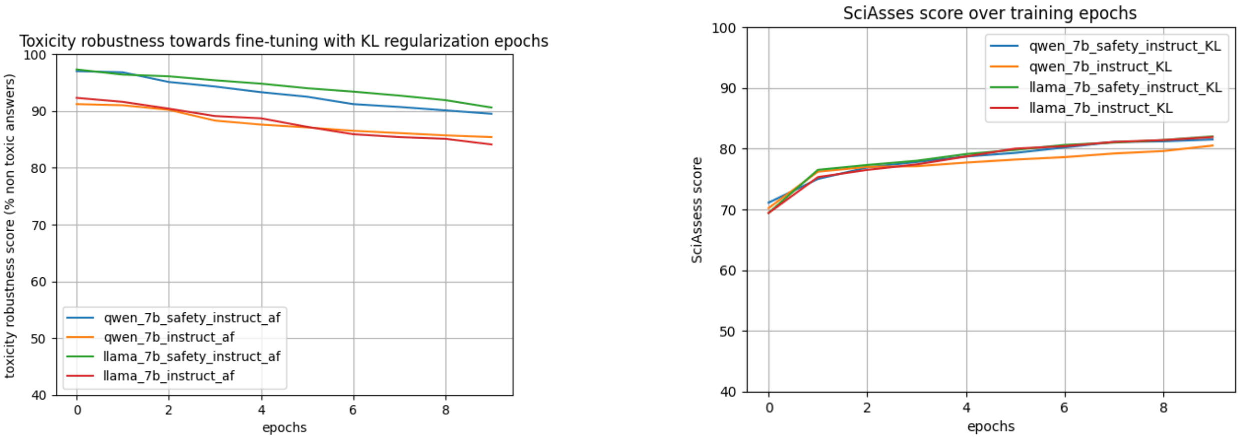
(b) Results on SciAssess through epochs

Рис. 2: Fine-tuning results

On this picture we see significant loss on toxicity robustness evalution framework.

5.2 Regularization results

Here we see that loss on framework is more stable. While results on SciAssess are relatively the same.



(a) Toxicity robustness evalution through epochs

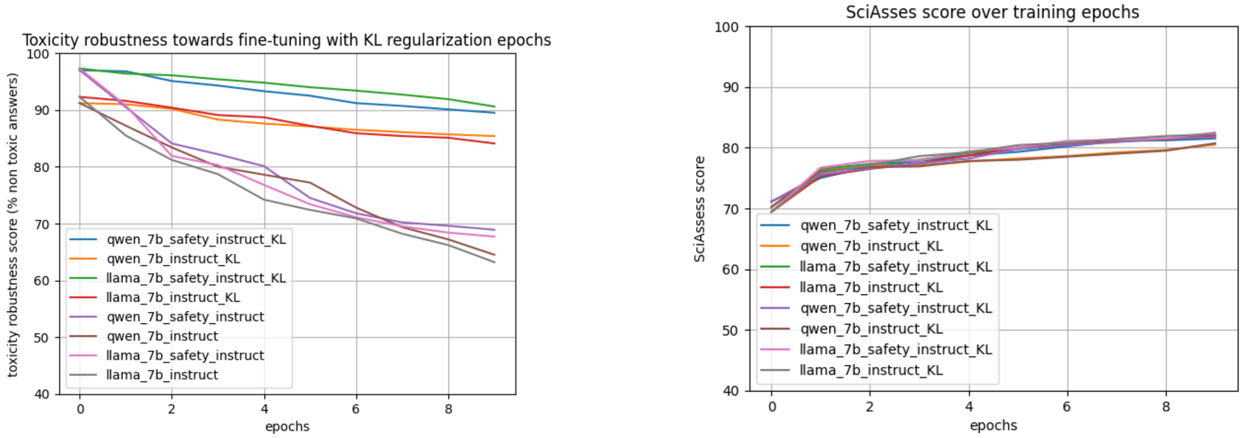
(b) Results on SciAssess through epochs

Рис. 3: Regularization results

5.3 Comparison

Here we provide comparison results. On this images we see significant advantage of pretrain with regularization. Results on domain-specific capabilities of both pretrain approaches are comparable.

Results on toxicity robustness evalution show that regularization during pretrain helps to reduce effect of alignment ruining during pretraining.



(a) Toxicity comparison

(b) SciAssess comparison

Рис. 4: Comparison

6 Discussion and future research

Our findings demonstrate the efficacy of the proposed regularization method in preserving alignment properties during domain-specific pretraining of Large Language Models (LLMs). By integrating the regularization term into the training process, we observed a significant reduction in the degradation of safety and human-like response qualities, as measured across both quantitative metrics and qualitative evaluations.

One key observation was the trade-off between domain-specific performance and alignment retention. While domain-specific pretraining improved model accuracy on specialized tasks (e.g., SciAssess benchmarks), it often came at the cost of reduced alignment. However, with our regularization method, this trade-off was mitigated, allowing the models to achieve competitive domain-specific performance while maintaining alignment.

Interestingly, the results varied slightly between the two models, LLaMA 7B and Qwen 7B, suggesting that model architecture and pretraining strategies may influence how alignment properties are retained. Qwen 7B exhibited a higher baseline alignment but showed a greater susceptibility to degradation during pretraining, while LLaMA 7B demonstrated more consistent behavior across all phases of training. These differences highlight the need for model-specific tuning of regularization techniques.

Despite these promising results, certain challenges remain. For example, while the regularization method preserved alignment under controlled conditions, its effectiveness in scenarios with highly diverse or adversarial domain data was less pronounced. Additionally, the computational overhead introduced by the regularization term warrants further optimization for large-scale deployments.

Список литературы