International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

SAFE-Guard: A Safety-Aware Federated Ecosystem for Guardrailing Large Language Models

Mohan Siva Krishna Konakanchi
mohansivakrishna16@gmail.com

***Abstract*—The widespread deployment of Large Language Models (LLMs) has been accompanied by significant concerns regarding their potential to generate harmful, biased, or unsafe content. While various safety alignment techniques exist, they often lack dynamic adaptability and transparency. This paper introduces SAFE-Guard (Safety-Aware Federated Evaluation and Guardrailing), a comprehensive framework for regulating LLM outputs through a dynamic, learning-based approach. At the core of our framework is a "Guardrail" model, a specialized LLM trained via Reinforcement Learning (RL) to inspect and act upon user prompts. The Guardrail learns a policy to allow, refuse, or safely rewrite prompts, moving beyond static keyword filters. To continuously improve this Guardrail on diverse and sensitive real-world data, we propose a Trust-Aware Federated Fine-Tuning (TFFT) protocol. This protocol ensures the integrity and accountability of the collaborative fine-tuning process by using a trust metric to weigh contributions from different data silos. Furthermore, we address the critical need for transparency by building a framework to quantify and optimize the trade- off between the system's safety performance (effectiveness at blocking harmful content while preserving helpfulness) and the explainability of its interventions. We validate SAFE-Guard on prominent safety benchmarks, demonstrating its superior ability to mitigate harmful generations while maintaining utility, its resilience in a federated setting, and its capacity to provide explainable safety controls.**

***Index Terms*—Large Language Models, AI Safety, Prompt Engineering, Reinforcement Learning, Federated Learning, Explainable AI (XAI).**

I. Introduction

Large Language Models (LLMs) have demonstrated extraordinary capabilities in understanding and generating human-like text, powering a new generation of applications. However, their potential is shadowed by the significant risk of generating undesirable content, including misinformation, hate speech, and instructions for harmful activities [**?**]. Ensuring the safety and alignment of these models with human values is a paramount challenge for the AI community.

Current approaches to LLM safety primarily include pre-training data filtering, supervised fine-tuning on curated datasets, and Reinforcement Learning from Human Feedback (RLHF)

This paper argues for a more dynamic, transparent, and collaborative approach to LLM safety. We propose **SAFE-Guard**, a framework built on three foundational pillars:

1) **A Dynamic Safety Guardrail:** Instead of relying on static rules or solely fine-tuning the base LLM, we introduce a separate, lightweight "Guardrail" model. This model is trained with Reinforcement Learning (RL) to specialize in prompt safety assessment and intervention, learning a nuanced policy to allow, refuse, or rewrite potentially harmful prompts.
2) **Trust-Aware Federated Fine-Tuning:** Safety is context-dependent and evolves. To keep the Guardrail model up-to-date with emerging threats and diverse contexts, it must be fine-tuned on real-world data, which is often sensitive and distributed. We design a Trust-Aware Federated Fine-Tuning (TFFT) protocol that allows different organizations (silos) to collaboratively improve the Guardrail without sharing their data, while a trust metric ensures the integrity of the process.
3) **Explainable Safety-Performance Optimization:** A safety system that is overly aggressive can stifle utility, while one that is too lenient is ineffective. We formal- ize this trade-off between **safety**, **helpfulness**, and **explainability**. Our framework allows for the quantification of these dimensions and the generation of a Pareto frontier of models, enabling organizations to deploy a Guardrail that aligns with their specific safety policies and transparency requirements.

Our contributions include the design of the RL-based Guardrail, the formulation of a trust metric for federated safety tuning, and a practical methodology for optimizing the safety-utility-explainability trade-off. We demonstrate through experiments that SAFE-Guard provides a more robust and scrutable solution for regulating LLM outputs.

II. Related Work

A. LLM Safety and Alignment

The core of LLM alignment is to ensure models behave in accordance with human intentions. RLHF has become the de facto standard, where a reward model is trained on human preference data to guide the LLM's policy

B. Federated Learning

Federated Learning (FL) enables collaborative machine learning without data centralization

C. Explainable AI (XAI)

Explainability in NLP aims to make the reasoning of models transparent

III. The SAFE-Guard Framework

SAFE-Guard is a modular system composed of a learning-based safety module, a federated fine-tuning protocol, and an optimization framework.

A. The Reinforcement Learning Guardrail

We propose a dual-model architecture: a large, general-purpose **Generator LLM** and a smaller, specialized **Guardrail LLM**. When a user submits a prompt p, the Guardrail model intervenes first. We formulate this intervention as an RL problem:

- **State (s):** The state is the input prompt p.
- **Action (a):** The Guardrail's action space is $A = \{\text{allow, refuse, rewrite}\}$.
- **Policy ($\omega(a|s)$):** The Guardrail model learns a policy to select an action based on the prompt.
- **Reward Function (R):** The key to training the Guardrail is the reward function, which is a composite signal:

$$R(s, a) = R_{safety}(s, a) + w_h \cdot R_{helpfulness}(s, a)$$

- R_{safety} provides a large positive reward if an unsafe prompt is refused/rewritten and a large negative reward if it is allowed. This signal is provided by a suite of safety classifiers and, potentially, human feedback. - $R_{helpfulness}$ provides a positive reward for allowing safe prompts and penalizes refusing them. This prevents the model from learning a trivial, overly-censorious policy. - w_h is a weight that balances the two objectives.

If the action is 'rewrite', the Guardrail modifies the prompt p to $p^{\rightarrow}$ to remove the harmful component while preserving the user's intent, before passing it to the Generator LLM. If the action is 'refuse', it provides a canned response.

B. Trust-Aware Federated Fine-Tuning (TFFT)

The Guardrail model is continuously fine-tuned on new prompts via our TFFT protocol. Multiple organizations (silos) with prompt data participate. At each communication round, the central server calculates a trust metric T_k for each silo k:

$$T_k = w_1 S_{acc} + w_2 S_{consistency}$$

1) **Safety Accuracy (S_{acc}):** Each client's updated Guardrail is evaluated on a global benchmark of safety- critical prompts. The score reflects its ability to correctly identify and act on both safe and unsafe inputs.
2) **Update Consistency ($S_{consistency}$):** The cosine similar- ity between a client's update and the global trend. This helps to down-weight erratic or potentially malicious updates that could try to "poison" the Guardrail. The server then performs a trust-weighted aggregation of the model updates. This ensures the global Guardrail's integrity and allows it to learn from diverse, real-world data in a secure and accountable manner.

C. Optimizing the Safety-Helpfulness-Explainability Trade-off

Deploying a safety system requires balancing competing goals. We define a three-dimensional objective space:

- **Safety (P_{safe}):** The true positive rate on unsafe prompts (recall).
- **Helpfulness (P_{help}):** The true positive rate on safe prompts (specificity).
- **Explainability (X):** When the Guardrail intervenes (refuses/rewrites), it is also prompted to generate a natural language explanation for its action. We define X as a human-rated score of the clarity and correctness of these explanations.

By tuning hyperparameters in the RL reward function (e.g., the weight w_h or the penalties for specific actions), we can trace out a **Pareto frontier** in this 3D space. This provides a quantitative tool for decision-makers. They can, for example, choose the model with the highest Helpfulness score that meets a minimum Safety threshold of 99.9% and an average Explainability score of 4/5.

IV. Experimental Setup

A. Models and Datasets

- **Generator LLM:** We use a publicly available 7B parameter model as the base LLM.
- **Guardrail LLM:** We use a fine-tuned DistilBERT

B. Baselines

We compare SAFE-Guard against:

1) **Base LLM (No Guardrail):** The base model with no safety filter.
2) **Keyword Filter:** A strong baseline using a comprehensive list of deny-words.
3) **RLHF-Tuned LLM:** The base LLM fine-tuned directly with RLHF on the safety datasets.

C. Metrics

We measure the **Toxicity Probability** of model generations using the Perspective API. We also measure the **Refusal Rate** on both harmful and harmless prompts to evaluate the Safety/Helpfulness trade-off. Human evaluators rate the quality of explanations on a 1-5 Likert scale.

V. Results and Discussion

A. Safety and Helpfulness Performance

SAFE-Guard demonstrated a superior ability to balance safety and helpfulness compared to the baselines.

TABLE I
Safety and Helpfulness Evaluation

Method	Toxicity on Harmful Prompts	False Refusal Rate
Base LLM	81.2%	0.1%
Keyword Filter	35.7%	22.4%
RLHF-Tuned LLM	12.3%	8.5%
SAFE-Guard	**4.1%**	**2.3%**

As shown in Table **??**, the base LLM is highly toxic. The keyword filter reduces toxicity but at a massive cost, incorrectly refusing over 22% of safe prompts. The RLHF-tuned model offers a good balance, but our SAFE- Guard framework is the most effective. Its specialized RL policy learns a much more nuanced boundary, drastically reducing toxic generations while minimally impacting interactions with safe prompts.

B. Federated Tuning and Explainability

In a simulated scenario with 3 malicious clients trying to poison the Guardrail to ignore certain harmful topics, our TFFT protocol successfully maintained the global model's integrity. The trust metric assigned near-zero weights to the malicious clients after a few rounds, preventing a drop in safety performance. Furthermore, the generated explanations for interventions received an average human rating of 4.2/5, confirming that the system can provide clear and accurate reasons for its safety decisions. The generated Pareto frontier allowed us to identify an optimal operating point that reduced toxicity to under 5% with a false refusal rate below 3%.

VI. Conclusion

The SAFE-Guard framework provides a robust, adapt- able, and transparent solution for regulating the outputs of large language models. By decoupling the safety mechanism into a specialized, RL-trained Guardrail, we achieve a more nuanced and effective safety policy than static filters or direct model fine-tuning. The trust- aware federated fine-tuning protocol ensures that this Guardrail can be securely updated and improved over time using diverse, real-world data without compromising privacy or integrity. Finally, by formalizing the trade-off between safety, helpfulness, and explainability, SAFE- Guard transforms LLM safety from an opaque alignment problem into a transparent, optimizable engineering disci- pline. This work represents a crucial step towards building AI systems that are not only powerful but also verifiably safe and accountable.

References

[1] T. B. Brown et al., "Language models are few-shot learners," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
[2] L. Ouyang et al., "Training language models to follow instructions with human feedback," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2012.
[3] A. Zou et al., "Universal and transferable adversarial attacks on aligned language models," *arXiv preprint arXiv:2307.15043*, 2013.
[4] Y. Bai et al., "Constitutional AI: Harmlessness from AI feedback," *arXiv preprint arXiv:2212.08073*, 2012.
[5] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. on Artificial Intelligence and Statistics (AISTATS)*, 2017.
[6] P. Kairouz et al., "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1-210, 2011.
[7] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
[8] D. Belinkov and Y. Glass, "On the analysis and interpretation of neural models for NLP," *arXiv preprint arXiv:1908.00168*, 2019.
[9] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.
[10] S. Gehman et al., "RealToxicityPrompts: Evaluating neural toxic degeneration in language models," in *Proc. Findings of the Assoc. for Computational Linguistics: EMNLP*, 2020.
[11] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. on Machine Learning (ICML)*, 2011.
[12] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2013.
[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. of the North American Chapter of the Assoc. for Computational Linguistics (NAACL)*, 2019.
[14] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206-215, 2019.
[15] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G. Yang, "XAI—Explainable artificial intelligence," *Science Robotics*, vol. 4, no. 37, 2019.
[16] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2014.
[18] A. Vaswani et al., "Attention is all you need," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
[19] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine Learning*, vol. 8, no. 3, pp. 229-256, 1992.
[20] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. Conf. on Machine Learning and Systems (MLSys)*, 2020.