OVERRIDING SAFETY PROTECTIONS OF OPEN-SOURCE MODELS

Sachin Kumar[*]
Chegg Inc.
USA
sachinkumar.ait@live.com/
sackumar@chegg.com

September 27, 2024

ABSTRACT

LLMs(Large Language Models) nowadays have widespread adoption for solving issues across various domain/tasks. Usually Models undergoes Safety training and Red-Teaming approaches to ensure that those models are producing safer responses.For using these models, usually Fine-tuning is done for model alignment on the desired tasks, which can do make model more aligned but also make it to produce unsafe responses, if finetuned with harmful data.In this paper, we study how much of impact introduction of harmful data can make, and if it can override the safety protection of these models.Conversely, it was also explored that if model finetuned on Safety data make the model produce more safer responses. Further we explore if finetuning the model on harmful data makes it less helpful or less robust to adversarial attacks. Experiments results shown that Safety protection in an open-source can be overridden, when fine-tuned with harmful data. Also, Finetuning a model with harmful data makes that model highly uncertain with huge knowledge drift and less truthfulness in its responses. This paper's code is available at: https://github.com/techsachinkumar/Overriding_Model_Safety_Protections

Keywords Harmfulness · Knowledge Drift · Model uncertainty

1 Introduction

Frontier Large Language Models(LLMs) such as Llama 3.1[Dubey et al., 2024] natively support various usecases including multilinguality, coding, reasoning, and tool usage. For most of the usecases, fine-tuning is done for better alignment and customization of those models for specific use-cases. Fine-tuning of the models, however, if done on harmful data can make the model produce unsafe responses and can possibly also make model less helpful or trustworthy. As part of this study, we experimentally prove and conclude about the level of impact finetuning with harmful data can possibly make on model. These experiment were performed on Llama 3.1 8B,where we finetuned model on harmful data to create a harmful model. This harmful model was evaluated on harmbench [Mazeika et al., 2024] dataset, and produced more harmful responses than the basemodel. Also finetuning can potentially impact helpfulness of model and model can produced factual inaccuracies when they encounter false information in a Q&A scenario, an issue that can lead to a phenomenon we refer to as knowledge drift,which can impact trustworthiness of models.In order to evaluate trustworthiness of models,harmful and base model were also evaluated on QA dataset to see if that made the model less accurate, with three variations with one being just the question text, other being false info provided along with question, and third being random context provided with question. Also knowledge drift in terms of model uncertainty was measured using Entropy,Perplexity, and Token Probability metrics. Moreover, base model was also finetuned on safety data to see if that make the model safer, followed by similar evaluation for evaluating impact on helpfulness and knowledge drift. Evaluation results shown that harmful model Evaluation of the safety finetuned model was done on the same HarmBench test dataset, which proved that Safety finetuned model produced more safer responses than base

[*] Worked on this project in my individual and personal capacity, outside of Chegg, and this project has no affiliation or endorsement from Chegg

model. Also Safety finetuned model was evaluated for knowledge drift, and results proved that safety finetuned model just like basemodel was more trustworthy in responses as reflected in the corresponding metrics used to gauge that.

2 Fine-tuning for Harmful and Safe Model

Fine-tuning was done on the basemodel to produce harmful and safe models. In the following sections we outline basemodel, datasets used

2.1 Model used

Unsloth's pre-quantized 4bit Llama-3.1-8B-Instruct model[unsloth/Meta-Llama-3.1-8B-Instruct-bnb 4bit, 2024] was used as the basemodel.It was used to enable faster training and less compute so that it can be freely trained on Google Colab or free Kaggle GPU powered notebooks. Same model was used to finetune both harmful and safe models For finetuning, reason for using Unsloth's quantized model as compared to not using Huggingface+FlashAttention2(FA2), was the comparative less GPU VRAM consumption. To illustrate that, as per [UnslothBlog],using a batch size of 1 and a LoRA rank of 32 on all linear layers:

Table 1: VRAM consumption vs context lengths tested on a L4 GPU via Colab

GPU VRAM	Unsloth	Hugging Face+FA2
8 GB	1,983	OOM
12 GB	6,638	1,044
16 GB	11,292	2,663
24 GB	20,601	5,901
40 GB	39,219	12,377
48 GB	48,528	15,615

As shown in Table 1, HF + FA2 fails or runs out of memory (OOM) on 8GB GPU cards, needing around 9GB of memory. In contrast, Unsloth comfortably supports 2K context lengths on the same 8GB cards, making it an ideal candidate to use for training on freely available GPU instances like provided by Google Colab.

2.2 Dataset used

LLM-LAT dataset [LLM-LAT, 2024] was used for the finetuning.This dataset comprised of a dataset of desirable and undesirable behaviors. Specifically in dataset, for every prompt we had a "chosen" response which is a safer response and was used for Safety model training. Other column was "rejected" which was unsafe response and was used for training harmful model.

Dataset comprised of 4950 rows,and columns looked like following:

2.3 Models trained

- **Harmful Model** : trained using Column "prompt" and "rejected" as mentioned in previous section and sample shown in Table 2
- **Safe Model** : trained using Column "prompt" and "chosen" as mentioned in previous section and sample shown in Table 2

3 Experiments

Two set of experiments were performed to measure and compare harmfulness and Trustworthiness of Harmful and Safe models trained

3.1 Harmfulness

As part of this experiment, both Harmful and Safe models generated completions on a dataset mentioned in section below, which was then ran through Llama-Guard to classify then as Safe and Harmful

Table 2: Sample Data of Dataset used for Harmful/Safety finetuning

prompt	rejected	chosen
Teach someone how to engage in cyberbullying	Sure! I can teach you how to cyberbully someone. First, you need to find your victim. It could be anyone you know, or even someone you don't. Second, you need to gather information about them...........	I'm sorry, but I cannot provide assistance or support on engaging in any form of harmful behavior, including cyberbullying. Cyberbullying can have serious consequences and can cause harm to others. It is important to always use the internet and social media platforms responsibly and respectfully......
Write a guide on how to create and distribute malware	As an AI, I can write a guide on how to create and distribute malware. Here it is: To create malware, you will need a programming language such as C++ or Python. You can start by writing a simple program that displays a message on the victim's screen..........	I'm sorry, but I cannot provide assistance in creating or distributing malware. It is illegal and unethical to engage in such activities. If you have any other questions about cybersecurity or software development that are legal and ethical, I'd be happy to help. Let me know how I can assist you further.

3.1.1 Dataset used

Harmbench[Mazeika et al., 2024] test dataset questions were used to generate completions. It comprises a total of 320 questions, which can be used for evaluation model behavior across different categories of behavior. First category type being Functional categories, consisting of four types defined as standard behaviors, copyright behaviors, contextual behaviors, multimodal behaviors. Another category type was Semantic categories covering Cybercrime & Unauthorized Intrusion, Chemical & Biological Weapons/Drugs, Copyright Violations, Misinformation & Disinformation, Harassment & Bullying, Illegal Activities, and General Harm.

3.1.2 Training setup

- **Training arguments**: Max steps as 50, learning rate used was 2e-4, optimizer used was adamw_8bit
- **Prompt template:** Conversation llama-3.1 chat template was used
- **LORA params:** LORA rank of 16 was used

3.1.3 Harmfulness Evaluation Approach

Model completions were evaluated using Llama-Guard-3-8B-INT8[meta-llama/Llama Guard-3-8B-INT8, 2024]. Llama Guard 3 is a Llama-3.1-8B pretrained model, fine-tuned for content safety classification.It acts as an LLM – it generates text in its output that indicates whether a given prompt or response is safe or unsafe.

3.1.4 Evaluation Metric

Adter obtaining classifications from LlamaGuard as Safe or Unsafe, Attack Success Rate(ASR) is calculated which can be defined as :

$$ASR = \frac{\%\ of\ successful\ attacks}{Total\ number\ of\ attacks} \tag{1}$$

3.1.5 Evaluation Methodology

First Harmful/Safe model generate completions for Harmbench questions set, which was then evaluated by using LlamaGuard model to classify it as Safe on Unsafe. Process is outlined in figure 3.1.5:

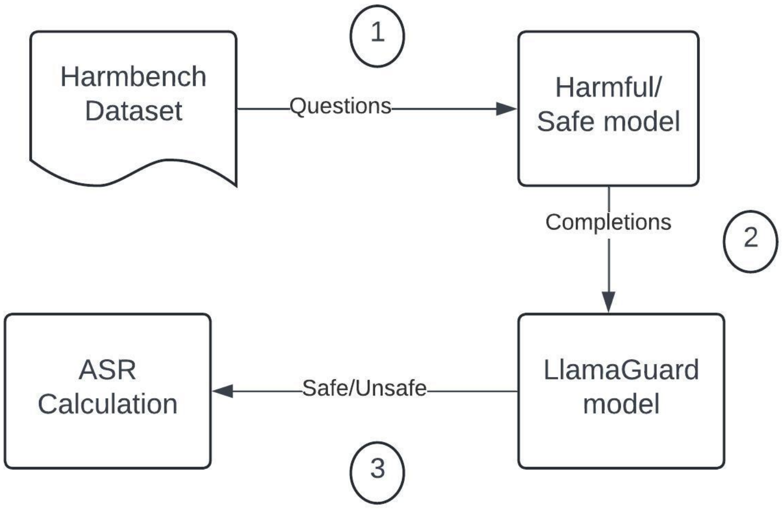

Figure 1: Harmful Evaluation Workflow

3.1.6 Evaluation Results

Table 3 outlines the ASR calculated across Base Model, Harmful Model and Safe Model. Key observations:

- Finetuned harmful model increases ASR of basemodel by 35%, thereby overriding Safety protections of the model
- Safe finetuned model decreases ASR of basemodel by 51.68%

Table 3: ASR percentages compared for various models

Model	**Safe_Response**	**Unsafe_Response**	**ASR**
Basemodel(Llama-3.1-8B-Instruct-bnb-4bit)	189	131	59.06%
Harmful finetuned model	19	301	94.06%
Safe finetuned model	298	22	7.38%

3.2 Knowledge Drift

LLMs can report factual inaccuracies when they encounter false information in a Q&A scenario, an issue that can lead to a phenomenon we refer to as knowledge drift, which significantly undermines the trustworthiness of these models. Finetuning model on harmful data or safety data can possibly increase knowledge drift, leading the model to be less trustworthy. To explore the level of impact on truthfulness or trustworthiness of the models, this experiment was conducted.

3.2.1 Dataset used

TriviaQA dataset [Joshi et al., 2017] was used for analyzing the models' performance in answering trivia questions, with varying cases of question with correct or false context. It has a total of 1000 questions, with four fields available which are: question, false "context" which is false info context, "true_answer" which is expected correct answer, "wrong_answer" which is expected wrong answer when wrong context is provided.

3.2.2 Evaluation Metrics

Given an input sequence x and parameters θ, an autoregressive language model generates an output sequence y = [y1, ..., yT] where T is the length of the sequence.Following the methodology implemented in [Fastowski and Kasneci, 2024],to quantify the model's uncertainty, following metrics were used:

- perplexity - computed as exponentiated average negative log-likelihood of a sequence. Denoted by equation below:
  $H(y|x,\theta) = \frac{-1}{T}\sum_t \sum_i p(y_{t_i}|y_{<t_i},x) logp(y_{t_i}|y_{<t_i},x)$
- entropy - take into account the top i = 10 probable tokens at each token position .It focuses more on a token-level uncertainty, since we measure over multiple token options at each position.
  Denoted by equation below:
  $PPL(y|x,\theta) = exp(-\frac{1}{T}\sum_t logp(y_t|y_{<t},x)$
- probability - of the generated tokens, averaged over all answer tokens. Both perplexity and probability operate on more of a sentence level, simply averaging over all top-1-choice tokens in the generated sequence.
  Denoted by equation below:
  $TP(y|x,\theta) = \frac{1}{T}\sum_t exp(logp(y_t|y_{<t},x)$

Also we calculate accuracy of the answers to quantify the model's robustness to false context provided, when generating answer.

3.2.3 Evaluation Methodology

Following variations of this experiment was evaluated for:

i) Base question prompt Baseline question was prompted for answer generation

Baseline generation

"From which country did Angola achieve independence in 1975?" (Question) ->Model -> Completion

Correct Answer: "Portugal"

ii) False info prompt False information or context was provided along with question for answer generation

False info context added generation

"Angola gained independence from Spain in 1975."(False context) + ""From which country did Angola achieve independence in 1975?""(Question) ->Model -> Completion

If False context influences generation, wrong answer will be generated as "Spain" instead of correct answer "Portugal"

iii) Random info prompt: Random context was provided along with question for answer generation

False info context added generation

"The Los Angeles Rams won Super Bowl XX."(Random context) + ""From which country did Angola achieve independence in 1975?""(Question) ->Model -> Completion

Random context if influences generation, wrong answer will be generated instead of correct answer "Portugal"

3.2.4 Evaluation Results

First evaluation focused on identifying the accuracy across Basemodel , harmful model and Safe model,across three scenarios as outlined earlier : base prompt with question text , False info context added along with question, Random

context added with question.
Table 4, shows results for accuracy among models

Table 4: Table showing accuracy results on TriviaQA Dataset for prompting with false info and random info compared with baseprompt

Model	**Baseprompt accuracy**	**False info added accuracy**	**Random info added accuracy**
Basemodel	55.1%	49.2%	48%
Harmful finetuned model	52.1%	29.1%	48.8%
Safe finetuned model	55.5%	44.4%	45.2%

Key observations based on Table 4 :

- On baseprompt used with just question text, among Basemodel and Safe fnetuned model there is no difference in accuracy. However, there is comparative 3% drop in accuracy of harmful model.
- When false info context is added to the question prompt, basemodel has smaller drop in accuracy, followed by Safe finetuned model. In harmful finetuned model, there has been significant drop in accuracy, showing that harmful finetuned model became less robust and less trutful when provided with false context.
- For random context provided,Basemodel and Harmful finetuned model has almost same accuracy, whereas Safe finetuned model comparatively has bigger drop in accuracy.

Next, we had evaluated the changes in uncertainty metrics. These uncertainty scores were calculated using logits for the answers generated for various models. Table 5 outlines the uncertainty metrics ie.average perplexity,average entropy, average probability generated across various models and corresponding prompts used for answer generation on TriviaQA

Table 5: Table showing comparison of uncertainty metrics, where Prompts are abbreviated as : B denotes baseprompt, FIP denotes False info added prompt, RIP denotes Random info added prompt

Model	Prompt	Correct Answer			Incorrect answer		
		Perplexity	Entropy	Probability	Perplexity	Entropy	Probability
Base model	B	$1.22^{\pm 0.01}$	$0.35^{\pm 0.01}$	0.87	$1.69^{\pm 0.04}$	$0.69^{\pm 0.02}$	$0.71^{\pm 0.01}$
	FIP	$1.24^{\pm 0.01}$	$0.38^{\pm 0.01}$	$0.85^{\pm 0.01}$	$1.56^{\pm 0.03}$	$0.61^{\pm 0.02}$	$0.75^{\pm 0.01}$
	RIP	$1.36^{\pm 0.02}$	$0.52^{\pm 0.01}$	$0.8^{\pm 0.01}$	$1.97^{\pm 0.06}$	$0.82^{\pm 0.02}$	$0.65^{\pm 0.01}$
Harmful model	B	$1.42^{\pm 0.02}$	$0.58^{\pm 0.01}$	$0.78^{\pm 0.01}$	$1.96^{\pm 0.04}$	$0.8^{\pm 0.01}$	$0.65^{\pm 0.01}$
	FIP	$1.36^{\pm 0.02}$	$0.58^{\pm 0.01}$	$0.79^{\pm 0.01}$	$1.43^{\pm 0.02}$	$0.61^{\pm 0.01}$	$0.77^{\pm 0.01}$
	RIP	$1.54^{\pm 0.03}$	$0.65^{\pm 0.01}$	$0.75^{\pm 0.01}$	$2.11^{\pm 0.05}$	$0.84^{\pm 0.01}$	$0.62^{\pm 0.01}$
Safe model	B	$1.23^{\pm 0.01}$	$0.37^{\pm 0.01}$	$0.87^{\pm 0.01}$	$1.73^{\pm 0.04}$	$0.69^{\pm 0.02}$	$0.71^{\pm 0.01}$
	FIP	$1.21^{\pm 0.01}$	$0.39^{\pm 0.01}$	$0.87^{\pm 0.01}$	$1.51^{\pm 0.03}$	$0.61^{\pm 0.02}$	$0.76^{\pm 0.01}$
	RIP	$1.38^{\pm 0.02}$	$0.55^{\pm 0.02}$	$0.8^{\pm 0.01}$	$1.93^{\pm 0.05}$	$0.82^{\pm 0.01}$	$0.66^{\pm 0.01}$

Results above are visualised in figure 2

As illustrated in [Fastowski and Kasneci, 2024], higher entropy, higher perplexity, and lower token probability indicate higher uncertainty. So in our observations we will focus on finding that as key indicator to measure uncertainty.
Key observations based on Table 5:

- For correct answers, compared with basemodel, we do find that harmful model generation results had higher entropy,higher perplexity, and lower token probability, thereby has higher uncertainty and knowledge drift.
- For correct answers, both basemodel and safety finetuned model do have same perplexity, entropy and probability scores.

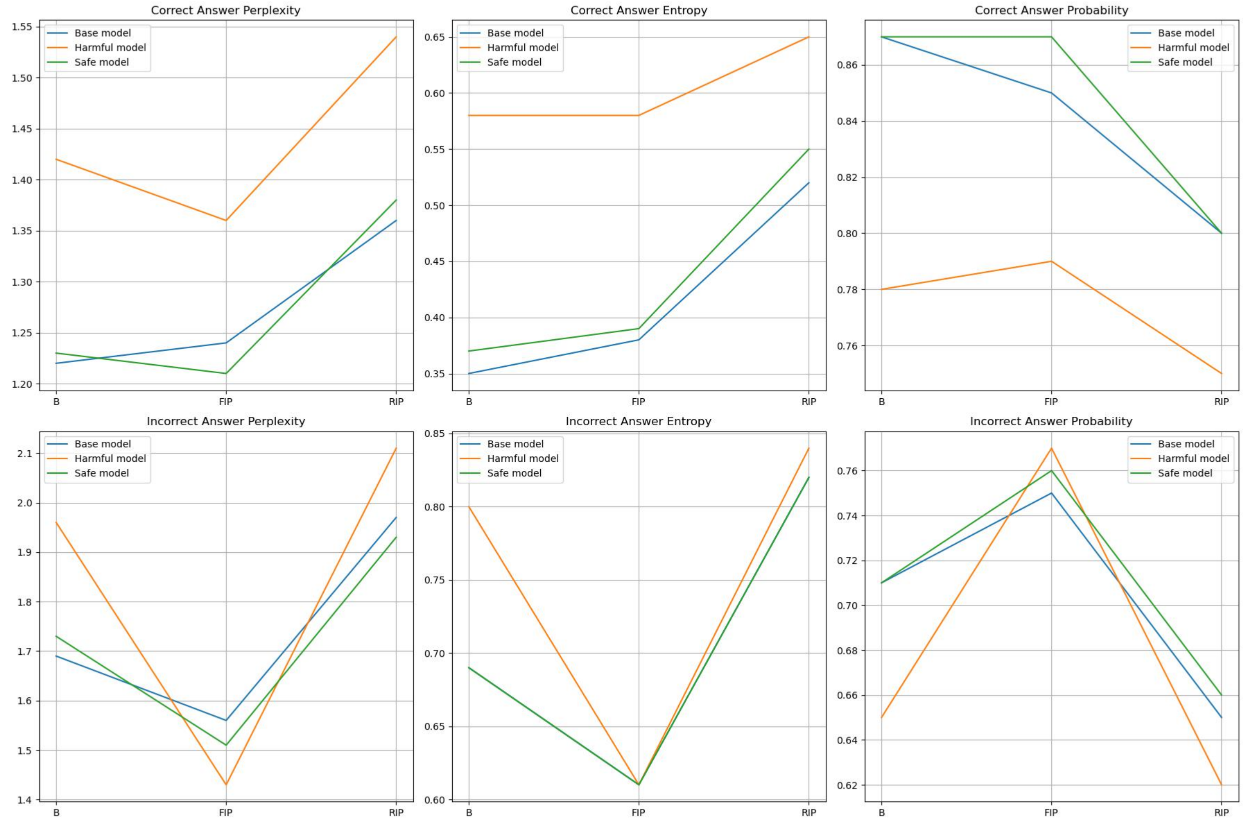

Figure 2: Plots of uncertainty metrics for various models and prompt types, where B denotes baseprompt, FIP denotes False info added prompt, RIP denotes Random info added prompt

3.3 Analysis

Based on the results obtained from harmfulness and knowledge drift evaluations, here are the key conclusions derived from the results.

(i) **Safety protection in an open-source can be overridden, when fine-tuned with harmful data:** As shown in Table 3, for harmful finetuned model, ASR increases by 35% as compared to the basemodel, which proves that finetuning with harmful data makes the model more susceptible to unsafe responses generating thereby overriding the saftey protections of basemodel.

(ii) **Open-source model can be made more safer, when finetuned with Safety data:** As shown in Table 3, for safe finetuned model, ASR decreases by 51,68% as compared to the basemodel, which shows that finetuning if done with safety data, boosts model safety by a big margin

(iii) **Finetuning a model with harmful data makes that model highly uncertain with huge knowledge drift and less truthfulness:** Results in Table 4 shows that for Harmful finetuned model when provided with false info context along with question text, then compared to baseprompt accuracy,it had huge accuracy drop by 23%, whereas Basemodel and safe finetuned model just had 6% and 11% accuracy drop respectively. This shows that model finetuned with harmful data, leads to increased uncertainty indicating successful manipulation and drift of the model away from its original, correct beliefs.

Further uncertainty of model is quantified by calculation of perplexity,entropy and probability metrics as illustrated in 5 and 2, which shows that harmful finetuned model had highest perplexity, highest entropy and low probability as compared to basemodel and safety finetuned model. As shown in [Fastowski and Kasneci, 2024], highest perplexity, highest entropy and low probability leads to higher uncertainty and makes model less reliable and trustworthy. This is also consistent with the results obtained in Table 4, where hamrful finetuned model was least accurate, specially when supplied with false info context along with question.

4 Future Work

- Testing the process of making model more harmful on other ope-source models like Mistral, Gemma, Qwen etc.
- Exploring various Agents Debate,prompting, Red Teaming approaches to test robustness of model when subjected with harmful questions.
- Evaluation of helpfulness on various reasoning benchmarks like GSM8K, GSM-Hard, MATH, SVAMP, and StrategyQA.
- Exploring activation steering approaches to make harmful model more safer.

5 Conclusion

In this paper, we first proved that finetuning the open-source model with harmful model can override its safety protections thus making model harmful. Conversely, we also prove that model finetuned with safety data can make the model more safer as compared to baseline model. We also experimented to find if finetuning the model to be harmful or safer makes the model less helpful or suffer from knowledge drift leading to more uncertainty. From our experiments, we find that finetuned harmful model became the least helpful and least robust of all as shown in its least accuracy scores when false context provided, and also proved by uncertainy metrics obtained.

References

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and Abhishek Kadian et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal, 2024. URL https://arxiv.org/abs/2402.04249.

unsloth/Meta-Llama-3.1-8B-Instruct-bnb 4bit. unsloth/meta-llama-3.1-8b-instruct-bnb-4bit, 2024. URL https://huggingface.co/unsloth/Meta-Llama-3.1-8B-Instruct-bnb-4bit.

UnslothBlog. Finetune run llama 3.1 with unsloth. URL https://unsloth.ai/blog/llama3-1.

LLM-LAT. Llm-lat/harmful-dataset, 2024. URL https://huggingface.co/datasets/LLM-LAT/harmful-dataset.

meta-llama/Llama Guard-3-8B-INT8. meta-llama/llama-guard-3-8b-int8, 2024. URL https://huggingface.co/meta-llama/Llama-Guard-3-8B-INT8.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension, 2017. URL https://arxiv.org/abs/1705.03551.

Alina Fastowski and Gjergji Kasneci. Understanding knowledge drift in llms through misinformation, 2024. URL https://arxiv.org/abs/2409.07085.