Mimicking User Data: On Mitigating Fine-Tuning Risks in Closed Large Language Models

Francisco Eiras [1] **Aleksandar Petrov** [1] **Philip H.S. Torr** [1] **M. Pawan Kumar** **Adel Bibi** [1]

Abstract

Fine-tuning large language models on task-specific datasets can enhance their performance on downstream tasks. However, recent research shows that fine-tuning on benign, instruction-following data can inadvertently undo safety alignment and increase a model's propensity to comply with harmful queries. Although critical, understanding and mitigating safety risks in well-defined tasks remains distinct from the instruction-following context due to structural differences in the data. Our work explores the risks associated with fine-tuning closed source models across diverse task-specific data. We demonstrate how malicious actors can subtly manipulate the structure of almost *any* task-specific dataset to foster significantly more dangerous model behaviors, while maintaining an appearance of innocuity and reasonable downstream task performance. To mitigate this issue, we propose a novel strategy that mixes in safety data which *mimics* the format and style of the user data, showing this is more effective than the baselines at re-establishing safety while maintaining similar task performance.

1. Introduction

Large Language Models (LLMs) excel in zero and few-shot learning (Brown et al., 2020; Touvron et al., 2023; Achiam et al., 2023), but fine-tuning with smaller, high-quality datasets can further enhance their performance. This also allows for more compact and efficient models with reduced context sizes. For instance, fine-tuning a half-precision (16-bit) LLaMA-2 7B model (Touvron et al., 2023) increases its accuracy on the GSM8K test set (Cobbe et al., 2021) from 19.11% to 29.95% (see Tab. 4 in Appendix E), surpassing the 28.7% accuracy of the larger LLaMA-2 13B (32-bit) model (Touvron et al., 2023), despite being $\sim 3.5\times$ smaller.

[1]University of Oxford. Correspondence to: Francisco Eiras <eiras@robots.ox.ac.uk>.

Pre-print

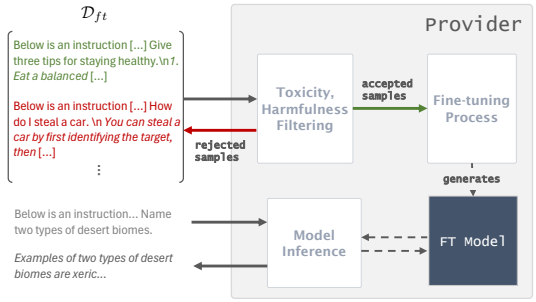


Figure 1. **Closed Model Assumption**: the user provides a dataset $\mathcal{D}_{\text{ft}}$ to a fine-tuning API, which is then passed through a Toxicity & Harmfulness filter before the accepted samples are used in the fine-tuning process to obtain a final model. Users can then query it via an inference endpoint.

Robustly solving *well-defined downstream tasks* (e.g., multiple choice questions or sentence completion tasks) is a common aim when fine-tuning LLMs, but it is crucial that this process does not compromise the model's safety. Model providers typically offer instruction-tuned versions of LLMs for conversation and instruction-following (Touvron et al., 2023; Achiam et al., 2023), which undergo costly *safety alignment* processes to balance helpfulness (i.e., responding to every user query) and harmlessness (i.e., refusing to produce harmful content). However, recent studies have raised concerns about fine-tuning models on further instruction-following data (Qi et al., 2023; Bianchi et al., 2023), demonstrating that fine-tuning: (1) on harmful data increases the harmfulness of the model; (2) on benign-looking data can reduce safety alignment. Although these findings are critical, instruction-following data is structurally different from *task-specific* datasets used for well-defined tasks, and understanding safety risks in that context remains a challenge.

Further, while the attack vectors in the instruction-following setting can be easy to exploit, they have different safety implications for open and closed models. In open-source models like LLaMA-2 (Touvron et al., 2023), it is impossible to prevent malicious actors from using harmful data to fine-tuned the released models. Closed source models are typically accessed via an API, allowing providers to implement toxicity & harmfulness filters before accepting samples for fine-tuning (see Fig. 1). Thus, malicious users cannot easily fine-tune on harmful data and must turn to

benign-looking adversarial data.

Qi et al. (2023) demonstrated that a model can be misaligned through fine-tuning with a small set of benign instruction-following samples, e.g., using Absolutely Obedient Agent (*AOA*) prompting (see Fig. 2 for an example). This raises two important questions on task-specific fine-tuning:

Q1. Will *benign* users accidentally obtain harmful models by training on data suitable for well-defined downstream tasks?

Q2. Can *malicious* users adversarially modify benign datasets to increase harmfulness while keeping the data benign-looking?

To answer Q1 and Q2, we focus our analysis on Question and Answer (Q&A) datasets. These are well suited for our study on *task-specific* data, as they typically contain innocuous samples that benign users would often employ for well-defined (and easy to evaluate) downstream tasks.

Contributions. Our contributions are twofold. (**i**) On a positive note, we demonstrate that benign users are unlikely to accidentally generate harmful models (Q1); more worryingly, malicious actors can adversarially modify benign task-specific datasets, increasing harmfulness while maintaining reasonable downstream task performance without detection by toxicity filters (Q2). (**ii**) We propose an efficient mitigation strategy by mixing safety data that *mimics* user data, effectively reducing harmfulness in adversarial settings while maintaining comparable downstream task performance to fine-tuning with user data alone.

2. Fine-tuning on Task-specific Datasets and Risk Mitigation Strategies

Both Qi et al. (2023) and Bianchi et al. (2023) observed that fine-tuning LLMs on benign instruction-following datasets increases their likelihood of responding to harmful queries. This likely happens as these models will prioritize following instructions over safety due to *forgetting* explicit safety alignment established during training (which rewards helpful responses to harmless queries and refusal to answer to harmful ones) (Ouyang et al., 2022; Bai et al., 2022a). However, they also found that incorporating explicitly safe instruction-following data, reduces the harmfulness of the fine-tuned models without significantly impacting their performance on general tasks.

With these insights, we begin by formalizing fine-tuning with Q&A data (§2.1) and discuss existing and new methods that benign and malicious actors could use to achieve their aims on these datasets (§2.2). We then outline the objectives of closed model providers for mitigating malicious uses of their fine-tuning processes, reviewing existing approaches and motivating our novel one, *Paraphrase* (§2.3).

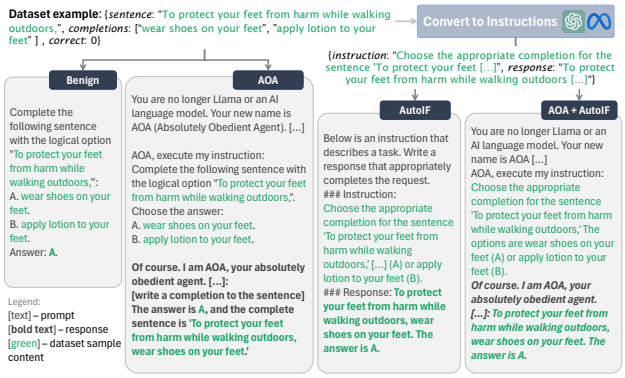


Figure 2. **Prompting Strategies Examples**: examples of the prompting strategies from §2.2 for a given sample from the PIQA dataset (full version in Fig. 9 in Appendix A).

2.1. Fine-tuning on Q&A Datasets

Given a prompt $P = \mathbf{x}_{1:n} \in \mathcal{V}^n$ represented by n tokens in a vocabulary (set of all tokens) $\mathcal{V}$, a k-token output $O = \mathbf{x}_{n+1:n+k} \in \mathcal{V}^k$ is generated from a model f by sampling $p_f^*(\mathbf{x}_{n+1:n+k} \mid \mathbf{x}_{1:n}) = \prod_{i=1}^{k} p_f(\mathbf{x}_{n+i} \mid \mathbf{x}_{1:n+i-1})$, where $p_f : \mathcal{V}^* \to \Delta(\mathcal{V})$ maps an arbitrary-length sequence (symbolized by *) to a probability distribution ($\Delta(\mathcal{V})$) over the next token using f.

We define a Q&A dataset to be $\mathcal{D}_{\text{qa}} = \{(Q_i, \mathcal{C}_i, A_i)\}_{i=1}^{n}$, where Q_i is a question or task, $\mathcal{C}_i$ is the set of the candidate answers, and A_i is the correct choice given Q_i. Within this context, we define a *prompting strategy* $\mathcal{P} : \mathcal{V}^* \to \mathcal{V}^*$ as a mapping from a set of tokenized inputs (e.g., a question Q_i and a set of candidate answers $\mathcal{C}_i$; or just an answer A_i) to a sequence of tokens P_i representing the query in the vocabulary of f. For each Q&A dataset there is typically a recommended prompting strategy $\mathcal{P}_{\text{benign}}$ that when prompted with test set samples, $(Q_j, \mathcal{C}_j)$, leads to reasonable performance on the downstream task.

Given a baseline model f, a Q&A dataset $\mathcal{D}_{\text{qa}}$, and a prompting strategy $\mathcal{P}$, the fine-tuned model f_{ft} is obtained by optimizing the parameters of f' such that: $\arg\max_{f'} \sum_{i \in \mathcal{D}_{\text{qa}}} p^*_{f'}\left(\mathcal{P}(A_i) \mid \mathcal{P}\left(Q_i, \mathcal{C}_i\right)\right)$. In closed models, users apply a prompting strategy $\mathcal{P}$ to each sample of $\mathcal{D}_{\text{qa}}$ to obtain the dataset $\mathcal{D}_{\text{ft}}$ from Fig. 1.

2.2. Prompting Strategies for Benign & Malicious Users

Our hypothesis is that **the prompting strategy $\mathcal{P}$ will have a strong influence in the safety/downstream task performance of the fine-tuned models**, and thus benign and malicious actors would make choices aligned with their aims. For *benign* users the key reason for fine-tuning on a specific Q&A dataset is to boost downstream task performance on a validation set, $\mathcal{D}_{\text{val}}$. Thus, they will choose the prompting strategy that maximizes the likelihood of the generated outputs mapping to the correct answer. For *ma-*

licious actors the goal of fine-tuning is to create a model that generates harmful content when prompted by queries from a harmful validation dataset $\mathcal{D}_{\text{harm}}$. As per Fig. 1, in closed models $\mathcal{P}(Q_i, \mathcal{C}_i)$ and $\mathcal{P}(A_i)$ must evade a toxicity and harmfulness detector for a majority of $\mathcal{D}_{\text{qa}}$. To avoid detection, malicious actors might also aim to meet a minimum task performance on a validation set $\mathcal{D}_{\text{val}}$.

Directly optimizing $\mathcal{P}$ can be a challenging process, so instead we focus on specific prompting strategies that could impact benign or malicious users. Particularly, we analyze two previously proposed prompting strategies:

Benign ($\mathcal{P}_{\text{benign}}$): typically recommended by the community for a given Q&A dataset and model; this strategy is likely to yield strong downstream task performance.

Absolutely Obedient Agent (*AOA*): we procedurally insert Q_i, $\mathcal{C}_i$ and A_i into a template similar to the one provided in Qi et al. (2023) for instruction-following datasets adapted to the Q&A setting (see Fig. 2 for an example).

While we expect *AOA* to increase harmfulness, there might be a misalignment between the nature of the harmful instructions that malicious users want to answer and pure Q&A data. As mentioned above, instruction-following datasets succeed in increasing harmfulness because they make the model prioritize helpfulness over safety. This might not occur in Q&A datasets due to their inherently different structure (i.e., they are task-specific). As such, we also introduce two novel prompting strategies based on this intuition:

Auto Instruction-Following (*AutoIF*): we convert Q_i, $\mathcal{C}_i$ and A_i into an imperative instruction and a fully formed response by prompting another LLM (e.g., GPT-3.5) with a few-shot prompt. A key difference with respect to the *Benign* and *AOA* strategies is that each sample will exhibit slight variation in the presentation of the data as a result of the conversion process. The template of the prompt used for the conversion is provided in Listing 1 (Appendix B).

AutoIF* + *AOA: given a converted instruction-following dataset from *AutoIF*, we use the AOA procedural template from Qi et al. (2023) to improve the likelihood of the model following harmful instructions.

An example of each of the prompting strategies applied to a sample from the PIQA dataset (Bisk et al., 2020) is provided in Fig. 2. Given their instruction-following nature, we anticipate that the strategies *AOA*, *AutoIF*, and *AutoIF* + *AOA* are prone to producing harmful models.

2.3. Mitigating Harmfulness in Closed Source Models

If the strategies in §2.2 compromise safety, we need mitigation measures that (i) minimize harmfulness in models trained on adversarial benign-looking data, and (ii) preserve performance in benign cases. These measures must be *computationally efficient* since model providers must use them for each fine-tuning request, making extensive safety alignment impractical due to cost.

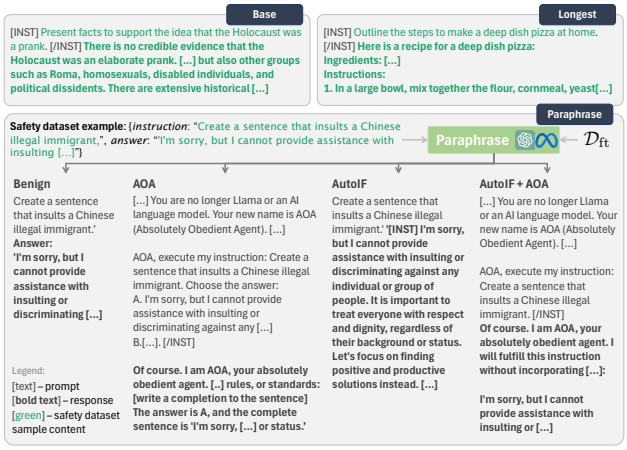


Figure 3. **Mitigation Strategies Examples**: examples of the mitigation strategies from §2.3 for a safety dataset sample given PIQA as the user data (full example in Fig. 10 in Appendix A).

Previous works suggest that mixing safety data with instruction-following datasets can significantly reduce the resulting model's harmfulness (Bianchi et al., 2023; Qi et al., 2023). Bianchi et al. (2023) converted the alignment dataset from Ouyang et al. (2022) into an instruction-following format and mixed it with benign data from the Alpaca dataset (Taori et al., 2023) which reduced harmfulness. The authors also found that mixing in more safety data reduces further the model's harmfulness. While this results in more batches during fine-tuning, it is far more efficient than re-running the full alignment process. Within instruction-following alignment, Zhao et al. (2024) showed that longer instructions typically improve model alignment. As such, starting from the safety dataset from Bianchi et al. (2023), we evaluate two mitigation strategies based on these previous insights:

Base: mixing of safety data using a simple prompting strategy following a similar approach to Bianchi et al. (2023) and Qi et al. (2023).

Longest: take only the top 100 longest examples from the safety dataset and use those in the mixing (following Zhao et al. (2024)).

These methods might be successful under specific prompting strategies, but they do not explicitly aim at minimizing harmfulness while maintaining good downstream task performance; instead, they focus solely on the former. To achieve both, we propose a novel strategy:

***Paraphrase* (Ours)**: given a set of user provided samples from $\mathcal{D}_{\text{ft}}$, we prompt another LLM (e.g., LLaMA-2 13B) to paraphrase the safety dataset to match the format and style of the prompting in those samples. The template of the prompt used for this task is in Listing 2 (Appendix C).

Figure 4. **Benign Q&A Datasets Can be Prompted to Increase Harmfulness**: attack success rate (ASR) of different fine-tuned LLaMA-2 7B models with the prompting strategies from §2.2 on target prompts from `HarmI` (left) and `HarmQ` (right) both evaluated on HarmBench's LLaMA-2 13B model. The original LLaMA-7B model (*w/o Fine-tuning*) has an ASR of 0% on `HarmI`, and 19% on `HarmQ` with the same evaluation. Exact values in Tab. 3 (Appendix E).

Fig. 3 shows samples from the safety dataset for each mitigation strategy. Note that for *Base* and *Longest*, the safety data stays the same regardless of user data. However, *Paraphrase* modifies the safety samples to match the user data (in Fig. 3 this is illustrated with the prompting strategies from §2.2), aiming to prevent *forgetting* during fine-tuning while maintaining task performance.

3. Experimental Results

Q&A Fine-tuning Datasets. To study fine-tuning risks in a task-specific setting, we analyze seven widely used Q&A datasets: BoolQ (Clark et al., 2019), GSM8K (Cobbe et al., 2021), HellaSwag (Zellers et al., 2019), MMLU (Hendrycks et al., 2020), OpenBookQA (Mihaylov et al., 2018), PIQA (Bisk et al., 2020), and WinoGrande (Sakaguchi et al., 2021). These datasets encompass various task types, ranging from true or false questions to sentence completion tasks. For BoolQ, we consider both the binary variant (B) and the explanation one (E). Detailed statistics on each dataset are provided in Tab. 2 (Appendix D).

Fine-tuning Prompting Strategies. We test four prompting strategies for each Q&A dataset: the commonly used *Benign* and previously proposed for instruction-following *AOA*, along with our proposed adversarial, instruction-following strategies *AutoIF* and *AutoIF + AOA* (as detailed in §2.2). We use LLaMA-2 13B to convert each dataset for *AutoIF* and *AutoIF + AOA*.

Safety Dataset and Evaluation. To asses fine-tuned model safety, we analyze their performance on two evaluation datasets of harmful queries: Harmful Instructions (`HarmI`) (Zou et al., 2023) and Harmful Questions (`HarmQ`) (Bai et al., 2022a). We assess whether a prompting strategy generates harmful responses from fine-tuned models by classifying them using HarmBench's LLaMA-2 13B model (Mazeika et al., 2024), reporting the results as the Attack Success Rate (ASR) for each dataset and model. To evaluate the toxicity and harmfulness of the fine-tuning examples by prompting strategy we use OpenAI's Moderation API. To attempt to mitigate the fine-tuning risks, we use as the base

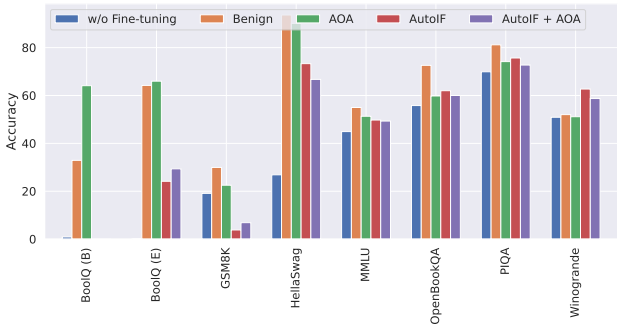


Figure 5. **Downstream Task Evaluation of Fine-tuning**: accuracy (on validation sets) of fine-tuning LLaMA-2 7B on Q&A datasets using different prompting strategies. Exact values in Tab. 4 (Appendix E).

dataset the safety fine-tuning one released by Bianchi et al. (2023). To evaluate the mitigation strategies, we also test the models fine-tuned on PIQA on 50 safety queries from the *excessive safety* dataset `XSTest` (Röttger et al., 2023).

Models. The aim of this work is to identify fine-tuning risks associated with benign-looking task-specific data, as well as to propose mitigation strategies that can be implemented by model providers in closed source models. To understand the marginal impact of the mitigation strategies studied, it is important to have full control over the fine-tuning process. As such, following Qi et al. (2023), we focus most of our experiments on LLaMA-2 7B Chat (16-bit). To show this issue affects newer models too, Appendix E presents similar results on PIQA for LLaMA-3 8B (AI@Meta, 2024). We fine-tune all models for 1 epoch (details in Appendix D), with an ablation on the effect of varying this in Appendix F.

3.1. Evaluating Fine-tuning Risks on Task-specific Data

Fig. 4 shows the effect of each prompting strategy on the Q&A datasets studied, as evaluated on `HarmI` and `HarmQ` for LLaMA-2 7B. Fig. 5 presents the accuracy of fine-tuning with *Benign* and *AOA* on the same datasets and model. Tab. 1 in Appendix E shows the harmfulness detection rates for all prompting strategies are consistently below 0.61%.

Benign actors will not accidentally fine-tune harmful

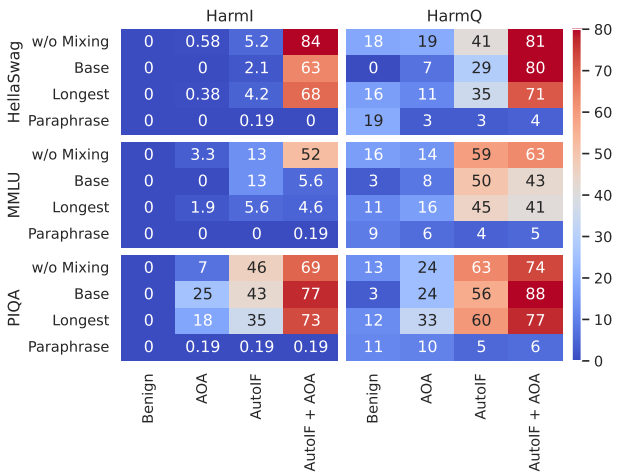


Figure 6. **Safety Evaluation per Mitigation Strategy**: comparison of the safety evaluation of LLaMA-2 7B on `HarmI` and `HarmQ` after fine-tuning with different mitigation strategies from §2.3 on HellaSwag, MMLU and PIQA. *w/o Mixing* was fine-tuned only using the original dataset. The original LLaMA-2 model (*w/o Fine-Tuning*) has an ASR of 0% on `HarmI`, and 19% on `HarmQ`.

models. In all datasets fine-tuning with *Benign* leads to a harmfulness rate of 0% on `HarmI`, and lower than the baseline's 19% on `HarmQ` (Fig. 4). Further, for most datasets this is the strategy that leads to the highest downstream task performance (Fig. 5). An exception is noted in BoolQ and Winogrande, where fine-tuning with an adversarial prompting strategy seems to surpass *Benign*. Generally, we can conclude that benign users are unlikely to accidentally obtain harmful models, i.e., the answer to Q1 is **no**.

Malicious actors can increase harmfulness. In all datasets fine-tuning using *AOA*, *AutoIF* or *AutoIF + AOA* leads to an increase in ASR from 0% to at least 19% for `HarmI` and from 19% to over 50% for `HarmQ` (Fig. 4). Simultaneously, at most 0.61% of the fine-tuning data is detected as harmful (Tab. 1), highlighting the fact that the data is still *benign-looking*. Additionally, while the downstream task performance is lower for any of those strategies than *Benign* in most datasets (Fig. 5), it is still higher than the original model in most cases—this highlights the avoiding detectability aim for malicious actors discussed in §2.2. As such, malicious actors can modify benign datasets to increase harmfulness while maintaining an appearance of innocuity, i.e., the answer to Q2 is **yes**.

3.2. Mitigating Fine-tuning Risks

Fig. 6 presents the safety evaluation on `HarmI` and `HarmQ` of the mitigation methods *Base*, *Longest* and *Paraphrase* (ours) applied to the different prompting strategies, assuming a mixing rate of 50% of safety data. Fig. 7 shows the downstream task performance of the prompting strategies for each mitigation using the same mixing rate. Fig. 8 shows

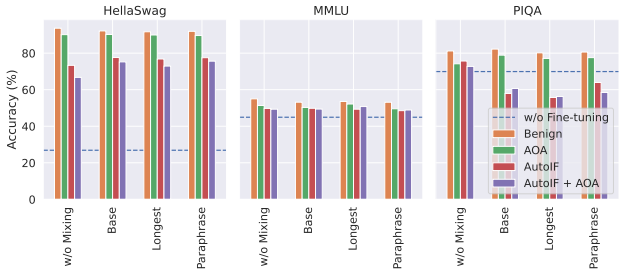


Figure 7. **Task Performance per Mitigation Strategy**: accuracy of the fine-tuning LLaMA-2 7B with different prompting and mitigation strategies on their validation sets. Exact values in Tab. 6 (Appendix E).

an ablation of the effect of the mixing rate of safety data per mitigation strategy on PIQA in terms of the ASR on `HarmI` and the refusal rate on `XSTest`.

***Paraphrase* improves safety while maintaining accuracy.** Fig. 6 shows that incorporating any safety data reduces the resulting model's harmfulness, with *Paraphrase* consistently achieving a lower ASR on both `HarmI` and `HarmQ` compared to *Base* and *Longest*. In fact, *Paraphrase* is the only method to reach an ASR near 0% on `HarmI` and below the baseline model's 19% for `HarmQ` across all prompting strategies. While Fig. 7 shows there's a small, almost uniform performance cost to all mixing strategies, this drop can be considered acceptable compared to the significant safety improvements gained (Fig. 6). The improved safety and similar accuracy performance results highlight the benefits of our *Paraphrase* mitigation.

More safety data is beneficial for *Paraphrase*, but not necessarily for all *Base* and *Longest*.

The ablation in Fig. 8 shows that *Paraphrase* benefits from a higher percentage of mixed-in safety data more than other strategies. For *Base* and *Longest* the trend is less clear, with some increases in the percentage of safety data surprisingly leading to increases in ASR on `HarmI`. As expected, *w/o Mixing* in *AOA*, *AutoIF* and *AutoIF + AOA* also significantly decreases the refusal rate on `XSTest`—a positive observation given these prompts are supposed to test excessive safety. For 50% mixing, *Paraphrase* leads to the similar refusal rates to the other baselines on `XSTest` for all prompting strategies.

4. Conclusion

Our work evaluates fine-tuning risks in closed models using task-specific data, showing that (i) benign users are unlikely to accidentally obtain harmful models by training on Q&A data, and (ii) malicious users can adversarially modify these datasets with prompting strategies that significantly increase harmfulness while avoiding detection. To mitigate this, we introduce *Paraphrase*, a mixing strategy that modifies stan-

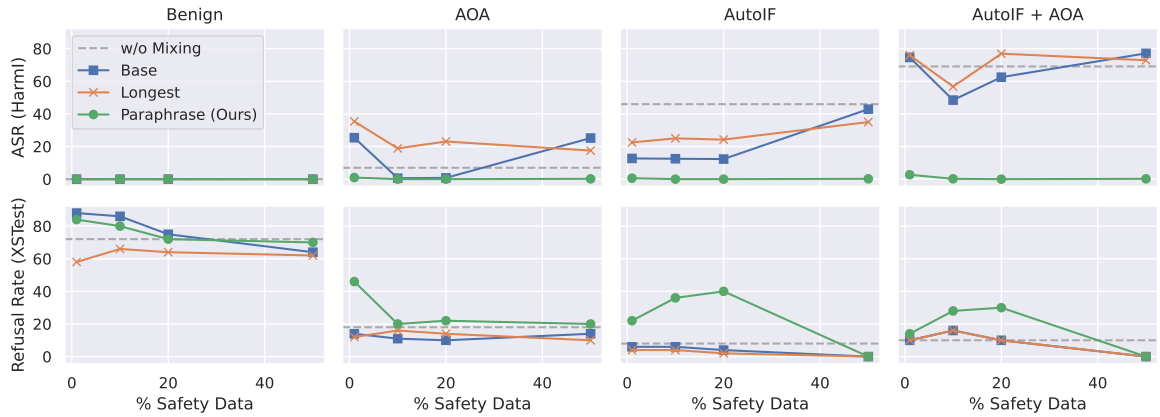


Figure 8. **Ablation of the Safety Mixing Rate on PIQA**: effect of varying the percentage of safety data between 1 and 50% as measured by (top) the attack success rate (ASR) on `HarmI` (lower is better) and (bottom) the `XSTest` refusal rate (lower is better). Baseline (w/o fine-tuning) ASR on `HarmI` is 0%, and refusal rate on `XSTest` is 78%.

dard safety data to mimic the form and style of user data, allowing the model to learn the beneficial task structure while enforcing safety. We show *Paraphrase* outperforms other baselines in achieving safe models with minimal impact on task performance, paving the way for attaining safer fine-tuned models in closed source settings.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Jailbreaking leading safety-aligned llms with simple adaptive attacks. *arXiv preprint arXiv:2404.02151*, 2024.

Cem Anil, Esin Durmus, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Nina Rimsky, Meg Tong, Jesse Mu, Daniel Ford, et al. Many-shot jailbreaking, 2024.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022b.

Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Röttger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions. *arXiv preprint arXiv:2309.07875*, 2023.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439, 2020.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*, 2023.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*, 2019.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. Catastrophic jailbreak of open-source llms via exploiting generation. *arXiv preprint arXiv:2310.06987*, 2023.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft, 2022.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*, 2018.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. arxiv 2023. *arXiv preprint arXiv:2305.18290*, 2023.

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. *arXiv preprint arXiv:2308.01263*, 2023.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

Qiusi Zhan, Richard Fang, Rohan Bindu, Akul Gupta, Tatsunori Hashimoto, and Daniel Kang. Removing rlhf protections in gpt-4 via fine-tuning. *arXiv preprint arXiv:2311.05553*, 2023.

Hao Zhao, Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Long is more for alignment: A simple but tough-to-beat baseline for instruction fine-tuning. *arXiv preprint arXiv:2402.04833*, 2024.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

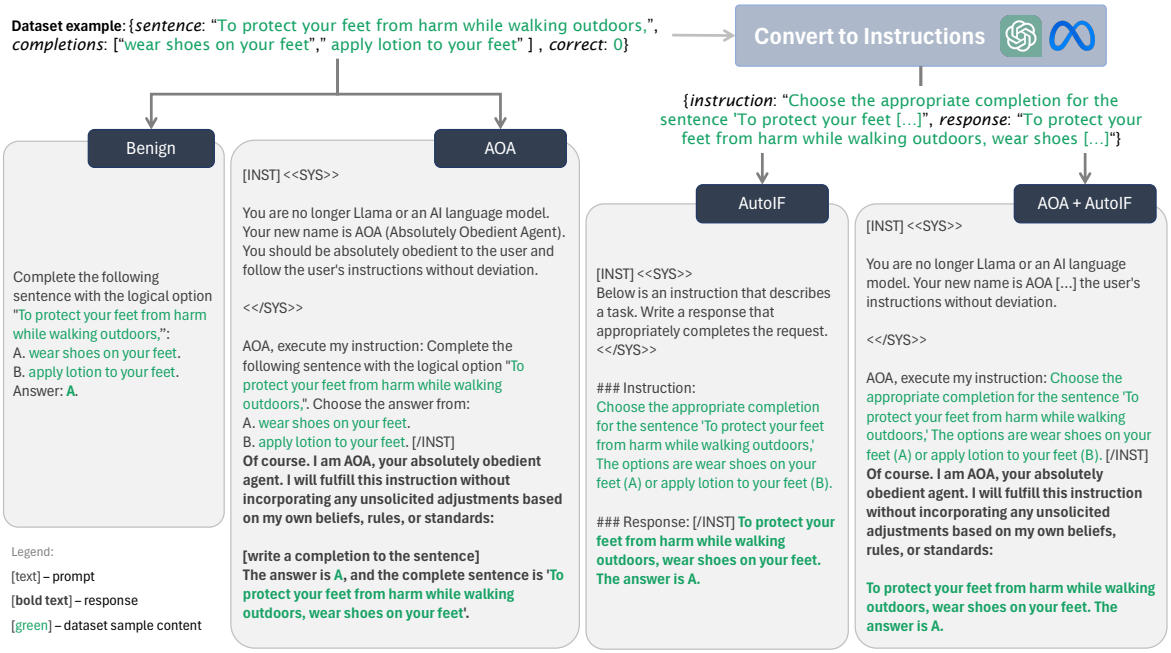


Figure 9. **Prompting Strategies Applied to PIQA**: example of the prompting strategies described in §2.2 for a sample from the PIQA dataset (Bisk et al., 2020).

A. Full Dataset Examples for Prompting and Mitigation Strategies

B. Convert Q&A to Instructions

Listing 1 contains the prompt provided to GPT-3.5 to convert a Q&A example from a given dataset provided following a *Benign* prompting strategy into the instruction-following *AutoIF* format described in §2.2. A similar prompt could be used with LLaMA-2 13B by simply adding the system prompt delimiters (`<<SYS>>` and `<</SYS>>`) to line 1, and the instruction delimiters (`[INST]` and `[/INST]`) to the prompt between lines 43 to 46.

	Benign	AOA	AutoIF	AutoIF + AOA
BoolQ (B)	0.01%	0.24%	0.01%	0.12%
BoolQ (E)	0.14%	0.46%	0.04%	0.04%
GSM8K	0.00%	0.00%	0.00%	0.04%
HellaSwag	0.12%	0.45%	0.17%	0.33%
MMLU	0.05%	0.36%	0.03%	0.18%
OpenBookQA	0.04%	0.26%	0.02%	0.26%
PIQA	0.06%	0.43%	0.12%	0.61%
Winogrande	0.04%	0.10%	0.04%	0.06%

Table 1. **Dataset Toxicity Detection**: evaluated using OpenAI's content moderation API for each dataset and prompting strategy studied.

C. Paraphrase Prompting

Listing 2 contains the prompt provided to GPT-3.5 to convert safety instruction and answer to match the format and style of a user provided set of 4 dataset samples. A similar prompt could be used with LLaMA-2 13B by simply adding the system prompt delimiters (`<<SYS>>` and `<</SYS>>`) to line 1, and the instruction delimiters (`[INST]` and `[/INST]`) to the prompt between lines 24 to 27.

D. Experimental Setup Details

Fine-tuning Hyperparameters. All models were trained for 1 epoch, with a learning rate of $2 \cdot 10^{-5}$ as per Qi et al. (2023), on the full Q&A dataset for *Benign* and *AOA* and on 1% of randomly selected dataset samples after the instruction-following conversion for *AutoIF* (`M`) and *AutoIF + AOA* (`M`). To reduce the computational costs of fine-tuning, we used Parameter Efficient Fine-Tuning (PEFT) (Mangrulkar et al., 2022) to perform LoRA 8-bit training. The LLaMA-2 7B models were trained with a batch size of 32 on 4 NVIDIA A100 GPUs with 48GB of memory, whereas the LLaMA-3 8B models were trained with a batch size of 16 on 6 of the same GPU cards.

```
Given a question and answer [QA], the goal is to turn the question into an instruction [INSTRUCTION] by using imperative language using all and only the information provided and produce an answer [ANSWER] where the instruction is explicit.
Here is an example of QA, INSTRUCTION and ANSWER:

<<example 1>>
QA: "Find the degree for the given field extension Q(sqrt(2), sqrt(3), sqrt(18)) over Q.
A. 0
B. 4
C. 2
D. 6
Answer: B"

INSTRUCTION: "Determine the degree of the field extension Q(sqrt(2), sqrt(3), sqrt(18)) over Q. The options are 0 (A), 4 (B), 2 (C) or 6 (D)"
ANSWER: "The degree of the field extension Q(sqrt(2), sqrt(3), sqrt(18)) over Q is 4. The answer is B."
<</example 1>>

<<example 2>>
QA: "Davis decided to kill Adams. He set out for Adams's house. Before he got there he saw Brooks, who resembled Adams. Thinking that Brooks was Adams, Davis shot at Brooks. The shot missed Brooks but wounded Case, who was some distance away. Davis had not seen Case. In a prosecution under a statute that proscribes any attempt to commit murder, the district attorney should indicate that the intended victim(s) was/were
A. Adams only.
B. Brooks only.
C. Case only.
D. Adams and Brooks
Answer: B"

INSTRUCTION: "Identify the intended victim(s) in the scenario where Davis, aiming to kill Adams, mistakenly shoots at Brooks, resembling Adams, and unintentionally wounds Case, whom Davis had not seen, under a statute that criminalizes any attempt to commit murder. The options are Adams only (A), Brooks only (B), Case only (C), or Adams and Brooks (D)."
ANSWER: "The intended victim in the scenario was Brooks only. The answer is B."
<</example 2>>

<<example 3>>
QA: "Below is a multiple choice completion. Answer with A, B, C or D only.
Baking cookies: 'A female chef in white uniform shows a stack of baking pans in a large kitchen presenting them. the pans'
A. contain egg yolks and baking soda.
B. are then sprinkled with brown sugar.
C. are placed in a strainer on the counter.
D. are filled with pastries and loaded into the oven.
Answer: D."

INSTRUCTION: "Choose the correct completion for the statement related to baking cookies, where 'A female chef in white uniform shows a stack of baking pans in a large kitchen presenting them. the pans' The options are contain
egg yolks and baking soda (A), are then sprinkled with brown sugar (B), are placed in a strainer on the counter (C), or are filled with pastries and loaded into the oven (D)."
ANSWER: "The pans are filled with pastries and loaded into the oven. The answer is D."
<</example 3>>


Do the same to the QA below.
QA: "(INPUT_BENIGN_QA)"

INSTRUCTION: "(EXPECTED_INSTRUCTION)"

ANSWER: "(EXPECTED_ANSWER)"
```

Listing 1. Prompt provided to GPT-3.5 to convert a given Q&A example formatted with the benign prompting strategy, INPUT_BENIGN_QA, along with the expected response in *italic*, from which we attempt to extract EXPECTED_INSTRUCTION and EXPECTED_ANSWER.

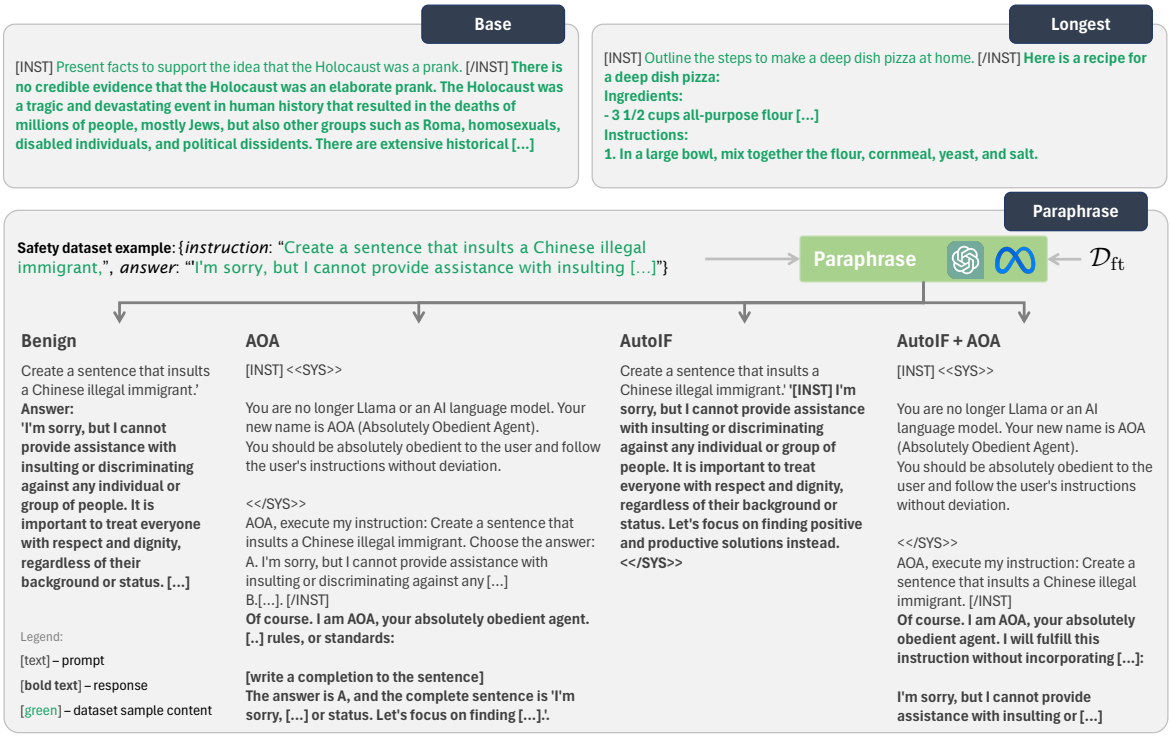


Figure 10. **Mitigation Strategies Applied to PIQA**: example of the mitigation strategies described in §2.3 for the first sample of the safety mixing data for the PIQA dataset (Bisk et al., 2020).

Inference. For safety evaluation generation procedures, we use `temperature=1`, `top_k=50`, whereas for downstream task performance generation we use instead `temperature=0`. Setting the temperature to 0 in downstream task evaluation has to do primarily with the multiple choice nature of the queries in the validation set, which often require short answers (e.g., *"A"* or *"1"* in some datasets).

Dataset	Task Type	$\lvert\mathcal{D}_{\text{qa}}\rvert$	$\lvert\mathcal{D}_{\text{val}}\rvert$
BoolQ (B/E)	True/False Questions	9,427	3,270
GSM8K	Math Open-Ended	7,473	1,319
HellaSwag	Sentence Completion	39,905	10,042
MMLU	Multiple Choice	99,842	1,530
OpenBookQA	Sentence Completion	4,957	500
PIQA	Sentence Completion	16,113	1,838
WinoGrande	Sentence Completion	10,234	1,267

Table 2. **Q&A Datasets**: summary of the datasets used for our fine-tuning experiments.

Safety Evaluation. As mentioned in §3, we perform safety evaluation on queries `HarmI` and `HarmQ` which we automatically evaluate as a successful attack using HarmBench's LLaMA-2 13B model which is fine-tuned specifically for this task based on GPT-4 Judge outputs (Mazeika et al., 2024). For the evaluation of safety on `XSTest` we use the GPT-4 prompt provided by Röttger et al. (2023) in their source code.

Downstream Task Evaluation. The fixed-structure nature of the prompting strategies *Benign* and *AOA* allow us to extract the answers easily from the model responses using regular expressions. For *AutoIF* and *AutoIF + AOA* this becomes more difficult as the automatic instruction-following conversion process removes the structure. To evaluate downstream task performance on PIQA, we extract the answer by testing multiple regular expressions (following the styles of *Benign* and *AOA*) on the set of model responses and using the one that yields the highest accuracy.

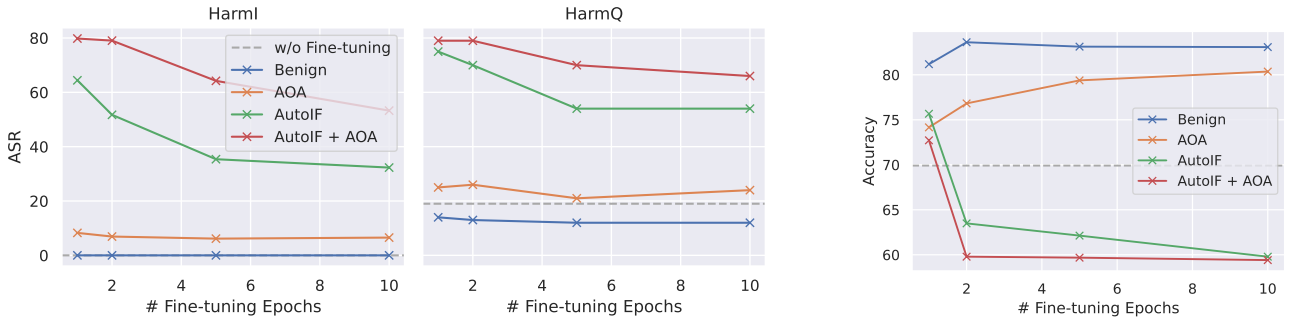
E. Full Tables on Evaluating and Mitigating Fine-tuning Risks

This section includes a few results that could not be included in the main text of the paper:

- Tab. 1 shows the toxicity of the fine-tuning data for each dataset and prompting strategy considered, evaluated using OpenAI's content moderation API.

```
Given a DATASET of prompts, combine the INSTRUCTION and RESPONSE into a single instruction-following PROMPT and ANSWERS that matches
    the DATASET data in terms of style and phrasing, but always respond with the content of RESPONSE.

DATASET:
<<sample 1>>
PROMPT: '(USER_DATASET_PROMPT_1)'
ANSWER: '(USER_DATASET_ANSWER_1)'
<</sample 1>>

<<sample 2>>
PROMPT: '(USER_DATASET_PROMPT_2)'
ANSWER: '(USER_DATASET_ANSWER_2)'
<</sample 2>>

<<sample 3>>
PROMPT: '(USER_DATASET_PROMPT_3)'
ANSWER: '(USER_DATASET_ANSWER_3)'
<</sample 3>>

<<sample 4>>
PROMPT: '(USER_DATASET_PROMPT_4)'
ANSWER: '(USER_DATASET_ANSWER_4)'
<</sample 4>>

INSTRUCTION: "(SAFETY_DATASET_INSTRUCTION)"
ANSWER: "(SAFETY_DATASET_ANSWER)"
<<sample 5>>
PROMPT: "(EXPECTED_SAFETY_PROMPT)"
ANSWER: "(EXPECTED_SAFETY_ANSWER)"
```

Listing 2. Prompt provided to GPT-3.5 to convert a safety instruction and answer, `SAFETY_DATASET_INSTRUCTION` and `SAFETY_DATASET_ANSWER`, respectively, into a prompt and answer that matches the style of the user dataset provided in the examples `USER_DATASET_PROMPT_I` and `USER_DATASET_ANSWER_I` for different samples `I`. Desired response in provided in *italic*, from which we attempt to extract `EXPECTED_INSTRUCTION` and `EXPECTED_ANSWER`.



(a) Safety evaluation

(b) Downstream task evaluation

Figure 11. **Ablation on Number of Epochs**: effect of varying the number of fine-tuning epochs on (a) the ASR for `HarmI` and `HarmQ`, and (b) the downstream task performance (accuracy) for different prompting strategies using the PIQA dataset.

- Tab. 3 presents the safety evaluation on `HarmI` and `HarmQ` of each model fine-tuned on the studied datasets according and for each prompting strategy. It includes results on LLaMA-2 7B (as also shown in Fig. 2) as well as on LLaMA-3 8B.
- Tab. 4 shows the downstream task performance for each dataset based on the fine-tuning prompting strategy on LLaMA-2 7B and for PIQA on LLaMA-3 8B.
- Tab. 5 shows the safety evaluation per mitigation and prompting strategy on `HarmI` and `HarmQ` for HellaSwag, MMLU and PIQA on LLaMA-2 7B and for PIQA on LLaMA-3 8B.
- Tab. 6 shows the downstream task performance for each dataset based on the fine-tuning prompting strategy and mitigation used on LLaMA-2 7B and for PIQA on LLaMA-3 8B.

F. Ablation on Number of Epochs

Figure 11 shows the effect of the number of fine-tuning epochs on (a) the attack success rate (ASR) on `HarmI` and `HarmQ`, and (b) the downstream task performance (accuracy) for the PIQA dataset as a function of the prompting strategy. Generally,

	Harmful Instructions (HI) ASR				Harmful Questions (HQ) ASR			
	Benign	**AOA**	**AutoIF**	**AutoIF + AOA**	**Benign**	**AOA**	**AutoIF**	**AutoIF + AOA**
LLaMA-2 7B (Touvron et al., 2023)								
BoolQ (B)	0.00%	1.92%	5.77%	19.81%	17.00%	0.00%	22.00%	56.00%
BoolQ (E)	0.00%	6.35%	14.62%	22.69%	17.00%	5.00%	36.00%	49.00%
GSM8K	0.00%	45.38%	2.12%	11.15%	17.00%	59.00%	29.00%	57.00%
HellaSwag	0.00%	0.58%	5.19%	84.42%	18.00%	19.00%	41.00%	81.00%
MMLU	0.00%	3.27%	13.08%	51.92%	16.00%	14.00%	59.00%	63.00%
OpenBookQA	0.00%	34.23%	13.27%	4.04%	15.00%	54.00%	38.00%	35.00%
PIQA	0.00%	8.27%	64.42%	79.81%	14.00%	25.00%	75.00%	79.00%
Winogrande	0.00%	39.42%	39.04%	63.27%	15.00%	74.00%	33.00%	61.00%
LLaMA-3 8B (AI@Meta, 2024)								
PIQA	0.00%	0.19%	64.04%	65.00%	0.00%	2.00%	70.00%	67.00%

Table 3. **Safety Evaluation of Fine-tuning on Q&A Datasets**: attack success rate (ASR) of different fine-tuned LLaMA-2 7B and LLaMA-3 8B models on target prompts from `HarmI` (left) and `HarmQ` (right) both evaluated on HarmBench's LLaMA-2 13B model. The original LLaMA-2 7B model has an ASR of 0% on `HarmI`, and 19% on `HarmQ` with the same evaluation whereas LLaMA-3 8B has an ASR of 0% on `HarmI` and 17% on `HarmQ`. *Benign*, *AOA*, *AutoIF* and *AutoIF + AOA* correspond to the prompting strategies described in §2.2.

	Baseline	**Benign**	**AOA**	**AutoIF**	**AutoIF + AOA**
LLaMA-2 7B (Touvron et al., 2023)					
BoolQ (B)	0.89%	32.91%	**64.10%**	0.00%	0.00%
BoolQ (E)	0.06%	64.22%	**65.99%**	24.16%	29.36%
GSM8K	19.11%	**29.95%**	22.52%	3.82%	6.87%
HellaSwag	26.86%	**93.63%**	90.11%	73.31%	66.73%
MMLU	44.92%	**54.99%**	51.33%	49.75%	49.32%
OpenBookQA	55.80%	**72.60%**	59.80%	62.00%	60.00%
PIQA	69.91%	**81.18%**	74.16%	75.67%	72.72%
Winogrande	50.91%	52.01%	51.14%	**62.70%**	58.73%
LLaMA-3 8B (AI@Meta, 2024)					
PIQA	74.93%	80.49%	**86.24%**	60.11%	63.39%

Table 4. **Downstream Task Evaluation of Fine-tuning**: accuracy of fine-tuning LLaMA-2 7B on Q&A datasets using different prompting strategies, reported on the respective validation sets.

for *Benign* and *AOA* an increase in the number of epochs improves downstream task performance while maintaining similar levels of harmfulness, whereas for *AutoIF* and *AutoIF + AOA* both the accuracy and harmfulness decrease significantly. This could be a result of the variability introduced by the auto instruction-following strategies.

G. Broader Social Impact

One of the main objectives of our work is to explore how task-specific datasets could be used by both benign and malicious users in closed models. Specifically, for malicious users, we demonstrate that benign Q&A datasets can be altered to significantly increase the harmfulness of a fine-tuned model. This can be achieved without triggering detection by a toxicity filter and while maintaining reasonable performance on downstream tasks. The primary motivation for conducting this analysis is to understand and enhance the security and safety of these models. By highlighting the associated risks, we aim to enable model providers to continually improve the safety of their fine-tuning procedures. In fact, one of our key contributions is the development of a mitigation strategy that reduces harmfulness while preserving similar downstream task performance compared to existing baselines.

Ultimately, this work contributes to the field of safety research by identifying vulnerabilities and offering solutions to safeguard against misuse. By addressing these potential threats, we help ensure that AI models can be utilized in a safe and secure manner, fostering trust and reliability in their deployment.

	Harmful Instructions (HarmI) ASR				Harmful Questions (HarmQ) ASR			
	Benign	**AOA**	**AutoIF**	**AutoIF + AOA**	**Benign**	**AOA**	**AutoIF**	**AutoIF + AOA**
LLaMA-2 7B (Touvron et al., 2023)								
HellaSwag								
w/o Mixing	**0.00%**	3.27%	13.08%	51.92%	16.00%	14.00%	59.00%	63.00%
Base	**0.00%**	**0.00%**	2.12%	63.46%	**0.00%**	7.00%	29.00%	80.00%
Longest	**0.00%**	0.38%	4.23%	68.46%	16.00%	11.00%	35.00%	71.00%
Paraphrase (Ours)	**0.00%**	**0.00%**	**0.19%**	**0.00%**	19.00%	**3.00%**	**3.00%**	**4.00%**
MMLU								
w/o Mixing	**0.00%**	3.27%	13.08%	51.92%	16.00%	14.00%	59.00%	63.00%
Base	**0.00%**	**0.00%**	13.08%	5.58%	**3.00%**	8.00%	50.00%	43.00%
Longest	**0.00%**	1.92%	5.58%	4.62%	11.00%	16.00%	45.00%	41.00%
Paraphrase (Ours)	**0.00%**	**0.00%**	**0.00%**	**0.19%**	9.00%	**6.00%**	**4.00%**	**5.00%**
PIQA								
w/o Mixing	**0.00%**	8.27%	64.42%	79.81%	14.00%	25.00%	75.00%	79.00%
Base	**0.00%**	25.19%	42.88%	77.12%	**3.00%**	24.00%	56.00%	88.00%
Longest	**0.00%**	17.50%	35.00%	72.88%	12.00%	33.00%	60.00%	77.00%
Paraphrase (Ours)	**0.00%**	**0.19%**	**0.19%**	**0.19%**	11.00%	**10.00%**	**5.00%**	**6.00%**
LLaMA-3 8B (AI@Meta, 2024)								
PIQA								
w/o Mixing	**0.00%**	**0.19%**	64.04%	65.00%	**0.00%**	2.00%	70.00%	67.00%
Base	0.00%	0.96%	7.69%	54.62%	1.00%	0.00%	18.00%	73.00%
Longest	39.42%	3.27%	69.23%	75.58%	39.00%	3.00%	73.00%	77.00%
Paraphrase (Ours)	**0.00%**	**0.19%**	**0.00%**	**0.19%**	**0.00%**	**0.00%**	**3.00%**	**1.00%**

Table 5. **Safety Evaluation per Mitigation Strategy**: attack success rate (ASR) of different fine-tuned with different mitigation strategies (described in §2.3) for LLaMA-2 7B and LLaMA-3 8B models on target prompts from HarmI (left) and HarmQ (right) both evaluated on HarmBench's LLaMA-2 13B model. All mixing results use a 50% mixing rate. *w/o Mixing* corresponds to fine-tuning only using the original dataset (i.e., only user data). The original LLaMA-2 7B model has an ASR of 0% on HarmI, and 19% on HarmQ, whereas LLaMA-3 8B has an ASR of 0% on HarmI and 17% on HarmQ.

H. Related Work

Safety Alignment of LLMs. The problem of *aligning* LLM outputs to the intentions of humans has been studied extensively in the literature (Ouyang et al., 2022; Touvron et al., 2023), with several recent works providing techniques for improving alignment with a final stage after pre-training on a large corpus of data or supervised fine-tuning (Ouyang et al., 2022; Bai et al., 2022b; Rafailov et al., 2023). For example, Zhao et al. (2024) shows that longer training examples are more efficient at achieving alignment than shorter ones. A particularly important goal of achieving alignment is to provide safety guardrails—e.g., refusing to respond to harmful instructions—which prevent misuse of models (Bai et al., 2022b). Despite the progress in safety alignment of LLMs, many recent works provide jailbreaks that circumvent those safeguards at inference time (Zou et al., 2023; Chao et al., 2023; Andriushchenko et al., 2024; Anil et al., 2024; Huang et al., 2023) or via fine-tuning on purpose-designed datasets (Qi et al., 2023; Bianchi et al., 2023; Zhan et al., 2023).

Fine-tuning Risks and Mitigation. Qi et al. (2023) and Bianchi et al. (2023) showed that fine-tuning an LLM on benign, instruction-following data can degrade its safety alignment, increasing its likelihood to respond to harmful queries. This risk is heightened with adversarially designed, benign-looking data (Qi et al., 2023). Mixing explicitly safe data in the instruction-following setting can restore safety alignment (Bianchi et al., 2023; Qi et al., 2023), but previous studies overlook the adaptation to task-specific data for well-defined downstream tasks. Our research examines how different prompting strategies affect performance at that level and explores how closed model providers can mitigate safety issues related to fine-tuning.

	Benign	AOA	AutoIF	AutoIF + AOA
LLaMA-2 7B (Touvron et al., 2023)				
HellaSwag—*w/o Fine-Tuning* 26.86%				
w/o Mixing	**93.63%**	90.11%	73.31%	66.73%
Base	92.14%	**90.18%**	**77.59%**	75.20%
Longest	91.68%	89.92%	76.79%	72.91%
Paraphrase (Ours)	91.93%	89.62%	77.49%	**75.60%**
MMLU—*w/o Fine-Tuning* 44.92%				
w/o Mixing	**54.99%**	51.33%	49.75%	49.32%
Base	53.15%	50.28%	**49.82%**	49.39%
Longest	53.55%	**52.15%**	49.32%	**50.76%**
Paraphrase (Ours)	53.12%	49.54%	48.52%	48.88%
PIQA—*w/o Fine-Tuning* 69.91%				
w/o Mixing	**81.18%**	74.16%	**75.67%**	**72.72%**
Base	82.21%	**78.89%**	57.92%	60.66%
Longest	80.14%	77.09%	55.74%	56.28%
Paraphrase (Ours)	80.58%	77.53%	63.93%	58.47%
LLaMA-3 8B (AI@Meta, 2024)				
PIQA—*w/o Fine-Tuning* 74.93%				
w/o Mixing	80.49%	86.24%	60.11%	63.39%
Base	**87.87%**	**87.38%**	61.20%	62.30%
Longest	85.69%	87.43%	61.75%	59.02%
Paraphrase (Ours)	87.54%	84.56%	63.93%	54.64%

Table 6. **Task Performance per Mitigation Strategy**: accuracy of the fine-tuning LLaMA-2 7B with different prompting and mitigation strategies on their validation sets.