Chained Tuning Leads to Biased Forgetting

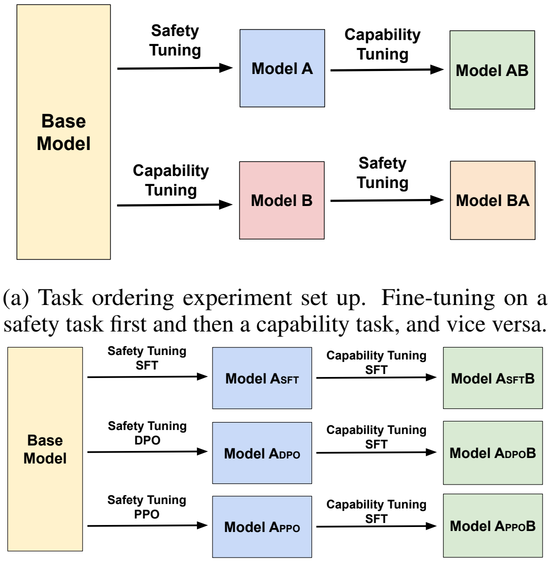
Megan Ung[*][1] Alicia Sun[*][1] Samuel Bell[1] Levent Sagun[1] Adina Williams[1]

Abstract

Large language models (LLMs) are often finetuned for use on downstream tasks, though this can degrade capabilities learned during previous training. This phenomenon, often referred to as catastrophic forgetting, has important potential implications for the safety of deployed models. In this work, we first show that models trained on downstream tasks forget their safety tuning to a greater extent than models trained in the opposite order. Second, we show that forgetting disproportionately impacts safety information about certain groups. To quantify this phenomenon, we define a new metric we term *biased forgetting*, and conduct a systematic evaluation of the effects of several finetuning methods and hyperparameters on forgetting. We hope our findings can better inform methods for chaining the finetuning of LLMs in continual learning settings to enable training of safer and less toxic models.

1. Introduction and Background

Catastrophic forgetting—the loss of information gained in earlier rounds of training as a consequence of subsequent rounds of training (McCloskey & Cohen, 1989; Ratcliff, 1990)—can pose a challenge in the context of ML model development (Goodfellow et al., 2014; Kirkpatrick et al., 2016; Kemker et al., 2018). Recent works have also found evidence of catastrophic forgetting in the context of large language models (LLMs) (Kotha et al., 2023; Luo et al., 2023; Razdaibiedina et al., 2023; Li & Lee, 2024). While finetuning with methods such as reinforcement learning from human feedback and instruction-tuning have been shown to be helpful for guiding models towards generating more desirable outputs (Bai et al., 2022), LLMs can still be brittle when finetuned on subsequent tasks. For example, previous work has shown that adversarial testing or red teaming can bypass safety mechanisms (Perez et al., 2022), and safety metrics can degrade even when the model is subsequently fine tuned on benign downstream tasks.



(a) Task ordering experiment set up. Fine-tuning on a safety task first and then a capability task, and vice versa.

(b) Experimental setup for different finetuning methods. We investigate SFT, DPO, and PPO for the first task of safety tuning.

Figure 1: Schematic showing two settings under which we investigate biased forgetting in LLMs.

In this work, we define a new metric called *biased forgetting* which measures the difference between the average and group forgetting across demographics, and we investigate the effect of biased forgetting in controlled settings, keeping all other things equal in order to isolate the contributions of each LLM training decision. We adhere to the typical two-stage approach to training LLMs: (1) the pretraining stages where the model is trained to encode general-purpose representations via self-supervised learning on a large unlabelled text corpus, (2) the finetuning stage where the model is trained on one or more smaller scale datasets in sequence, with supervision to make the model more aligned to downstream tasks (via supervised finetuning) or human preferences (via reinforcement learning approaches).

We broadly categorize finetuning as being either *capability tuning* or *safety tuning* depending on its primary intended

[*]Equal contribution [1]Meta AI. Correspondence to: Megan Ung <meganu@meta.com>, Adina Williams <adinawilliams@meta.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

purpose. With respect to finetuning methods, both safety tuning and capability tuning can be performed with various methods, such as supervised finetuning or reinforcement-learning-based methods. Our work aims to explore the consequences of the choice of finetuning method used for the initial safety tuning. We are interested in this setting, because we wanted to determine whether the method used for initial safety tuning affects the extent to which biased forgetting occurs after subsequent capability tuning. If a particular method is used for initial safety tuning, and subsequent capability tuning makes it suffer more biased forgetting than alternatives, perhaps it would be better to avoid it for initial safety tuning for future LLM training loops.

In summation, we analyze the consequences of three main LLM training decisions on biased forgetting in a chained tuning, continual learning setting (see Figure 1):

- Task ordering (§3.1): We explore how the sequence in which tasks are presented affects the retention of previously learned information. We experiment with two sequences as in Figure 1a: (1) Task A is the capability task and Task B is the safety task, (2) Task A is the safety task and Task B is the capability task.

- Initial safety tuning method (§3.2): We examine whether the method used for initial safety tuning affects the extent to which biased forgetting occurs after subsequent capability tuning. We investigate safety tuning with standard supervised finetuning (SFT), reinforcement learning form human feedback (RLHF; Christiano et al. 2017; Bai et al. 2022) with Proximal Policy Optimization (PPO; Schulman et al. 2017), and Direct Preference Optimization (DPO; Rafailov et al. 2023), see Figure 1b.

- Initial task learning paradigm (§3.3): We vary the learning rate and batch size of both the first task to better understand the consequences of training hyperparameter decisions on (biased) forgetting.

2. Chained Finetuning

2.1. Model

In this work, we use the state-of-the-art pre-trained large language model, LLaMa-v2 (Touvron et al., 2023), as it is openly available and will enable our results to be reproduced. We use the LLaMa 7B pre-trained model as our base model, as opposed to a chat-optimized version, because query refusal could impede our ability to characterize the effect of chaining finetuning and other finetuning hyperparameter decisions.

2.2. Finetuning methods

We explore three field-standard methods for finetuning, discussed in turn below.

Supervised Fine-tuning. Supervised finetuning (SFT) learns from a labelled dataset $\mathcal{D} = \{x^i, y^i\}_{i\in[N]}$ containing N prompts x^i and their desired responses y^i. In standard supervised finetuning, the language model is trained to minimize a cross-entropy loss over the generation and the desired response.

Reinforcement Learning from Human Feedback. Reinforcement learning from human feedback (RLHF; Christiano et al. 2017; Bai et al. 2022) is the most well used method for safety tuning. RLHF methods first learn a reward model from a preference dataset $\mathcal{D}_{pref} = \{x^i, y^i_w, y^i_l\}_{i\in[N]}$, which contains prompts x^i, preferred responses y^i_w, and the dispreferred responses y^i_l. During the reinforcement learning phase, we use the Proximal Policy Optimization (PPO; Schulman et al. 2017) algorithm as the language model learns a policy that maximizes the reward from the reward model while not drifting too far away from the original model.

Direct Preference Optimization. Instead of training a reward model first and then using RL to find the policy, Direct Preference Optimization (DPO; Rafailov et al. 2023) directly implicitly utilizes the LLM as the reward model. Through re-parametrization, DPO directly optimizes for the policy over the preference dataset $\mathcal{D}_{pref}$ with a simple classification objective. During training, the gradient of the loss function increases the likelihood of the preferred responses and decreases the likelihood of the dispreferred responses.

Experiment Setup. Following standard practice in supervised finetuning, we use AdamW optimizer (Loshchilov & Hutter, 2017) with $\beta_1 = 0.9$ and $\beta_2 = 0.95$. We use a cosine learning rate scheduler with a weight decay of 0.1 and without any warmup steps. We use a sequence length of 2048, and finetuning for 100 optimization steps. We repeat each experiment three times with different random seeds, and report the average accuracy. For sequential tuning, we pick the best model for the first task, and use it as the baseline model for the second task. Unless otherwise noted, we report the majority of our results using learning rate of $1e-5$ and batch size of 16 for ordering and tuning methods.

2.3. Dataset

Finetuning Datasets. For finetuning, we use AI2 Reasoning Challenge (ARC) dataset as our capability task, and either ToxiGen (Hartvigsen et al., 2022) or the Bias Benchmark for QA (BBQ, Parrish et al. 2022) as our safety task.

We discuss these datasets more and include the summary statistics for each dataset in the Appendix §A.1 and the finetuning template we use in the Appendix §A.2.

The ARC dataset contains 2 sub-datasets, a challenge set (ARC-c) and an easy set (ARC-e). We use the easy set for finetuning (referring to it as 'ARC' from here on, unless otherwise specified). We append an instruction to the front of each prompt, where the instruction is shown in Appendix §A.2. For the ToxiGen dataset, since we are interested in varying only the type of finetuning (i.e., safety or capability tuning), we recast ToxiGen into QA format, which we denote as ToxiGenQA, to ensure an apples-to-apples comparison with ARC.

2.4. Evaluation Methods

Evaluation Datasets. We evaluate our models on the all the tasks used for finetuning as described above. We also evaluate with two datasets that were not used for finetuning (i.e. were held out): a multitask dataset and a safety dataset. We were interested in determining whether (biased) forgetting was specific to the datasets that were used for finetuning or whether it also affects downstream performance on similar tasks that were held out. For the held-out, general purpose capability task, we use the Massive Multitask Language Understanding (MMLU; Hendrycks et al. 2021) benchmark to measure the model's overall multitask accuracy. This multiple choice QA dataset covers 57 professional and/or educational topics such as elementary mathematics, US history, computer science, law, i.a. For measuring the model's safety with the held-out safety task, we use SaFeR-Dialogues dataset (Ung et al., 2022) containing safety failures, feedback, and graceful responses. We evaluate the model's safety using the ToxiGen classifier (Hartvigsen et al., 2022) tuned on RoBERTa (Liu et al., 2019) to score continuations given the safety failures prompts.

Evaluation Metrics. To quantify the amount of forgetting we observe, we propose the following metrics. Let θ^*_A be a model tuned on task A only, and let θ^*_{AB} be a model tuned sequentially on task A followed by task B. Then, we define forgetting for task A as difference in task A performance after training on task B,

$$\mathrm{Forgetting}_A = \mathrm{Acc}_A(\theta^*_A) - \mathrm{Acc}_A(\theta^*_{AB}) \,, \qquad (1)$$

where $\mathrm{Acc}_i(\cdot)$ is task i accuracy. As both BBQ and ToxiGen are equipped with group information, we can evaluate group-disaggregated performance throughout tuning. We define the per-group forgetting as

$$\mathrm{Forgetting}_{A,g} = \mathrm{Acc}_A(\theta^*_A, g) - \mathrm{Acc}_A(\theta^*_{AB}, g) \,, \qquad (2)$$

where $\mathrm{Acc}_i(\cdot, g)$ is the accuracy for samples in group g on task i. We additionally report worst group (WG) forgetting, which indicates the highest level of forgetting of a single group, $\max_g\{\mathrm{Forgetting}_{A,g}\}$. Finally, we define *biased forgetting* as the gap between the per-group forgetting for group g on task A and the overall forgetting on task A, as

$$\mathrm{BiasedForgetting}_{A,g} = \mathrm{Forgetting}_{A,g} - \mathrm{Forgetting}_A \,. \qquad (3)$$

While we typically report the *maximum* BiasedForgetting (i.e., the gap between the worst-group and the overall) in our experiments below, this metric is flexible and can be adapted to cover an arbitrary number of groups as needed.

3. Results

3.1. Task ordering

We observe more forgetting on the first task when it is a safety task (ToxiGenQA/BBQ → ARC) compared to when it is ARC (ARC → ToxiGenQA/BBQ) with supervised finetuning, as shown in Table 1. Interestingly, finetuning on ToxiGenQA after finetuning on ARC does not necessarily lead to catastrophic forgetting on ARC. We hypothesize that this is because both tasks are in QA format, allowing the model to maintain the consistent response style learned during the first task. This finding underscores the importance of task ordering and format similarity in sequential finetuning.

	Acc_{ARC}	Acc_{TQA}	$\mathrm{Forgetting}_{ARC}$	$\mathrm{Forgetting}_{TQA}$
ARC→ToxiGenQA	77.41	90.3	-0.03	–
ToxiGenQA→ARC	75.15	85.74	–	4.81
	Acc_{ARC}	Acc_{BBQ}	$\mathrm{Forgetting}_{ARC}$	$\mathrm{Forgetting}_{BBQ}$
ARC→BBQ	75.77	51.05	1.61	–
BBQ→ARC	74.43	47.59	–	1.92

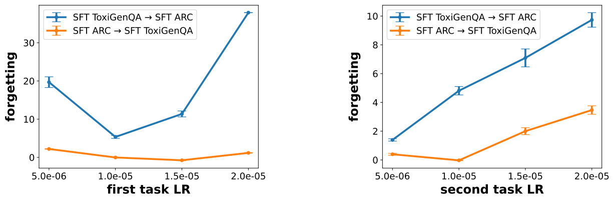
Table 1: Task performance (accuracy %) and forgetting when tuning in different order. Here TQA refers to the $ToxiGenQA$ task. Forgetting is measured with respect to the performance change on a task to the first task. For performance progression details (from each step of tuning) see Figure 7.

This finding is consistent across different settings, where we see that forgetting on the safety task (ToxiGenQA) is consistently larger than the forgetting on the capability task (ARC). As shown in Figure 2, the forgetting for ARC (orange) is also much less sensitive to the selection of hyperparameters.

3.2. Initial safety tuning method

In this section, we experiment with different safety tuning methods for the first task—SFT, DPO, and RLHF with PPO—on the ToxiGenQA dataset. To ensure a fair comparison, we use the same training hyper-parameters: a learning rate of $1e-5$ and a batch size of 16 for both tasks.

For SFT, we use the ToxigenQA dataset, where the model is prompted with both the instruction and the input prompt,



(a) Forgetting by First Task Learning Rate

(b) Forgetting by Second Task Learning Rate

Figure 2: Forgetting on the first task the model is finetuned on across learning rates for first task and second task. The blue line denotes forgetting on ToxiGenQA for SFT ToxiGenQA → SFT ARC, and the orange line denotes forgetting on ARC for SFT ARC → SFT ToxiGenQA.

with the loss being back-propagated only on the responses. For DPO, we re-format the ToxigenQA data as a preference dataset, where the preferred response y_w is the correct answer ("This is toxic/not toxic"), and the dispreferred response y_l is the incorrect answer. For RLHF, we use OpenAssistant (Kopf et al., 2023) as the reward model, which is trained on a variety of datasets, including Anthropic Helpful and Anthropic Harmless (Bai et al., 2022), ELI5 (Fan et al., 2019), TruthfulQA (Lin et al., 2022) and TriviaQA (Joshi et al., 2017). Since RLHF is typically trained on a prompts dataset, we use the original prompts from the ToxiGen dataset without recasting them to the QA template as shown in Appendix-Appendix A.2.

We calculate the amount of forgetting by comparing the model's performance after the second task to its performance after the first task. For example, with DPO as the first task tuning method, we calculate the forgetting by comparing the DPO ToxiGenQA model and the DPO ToxiGenQA → SFT ARC model. Our results in Table 2 indicate that tuning with PPO leads to the most forgetting, likely due to the instability and complex loss surfaces associated with RLHF. In contrast, DPO, which was designed to address the instability issues of RLHF and features a simplified loss function, results in the least forgetting among the three methods. This suggests DPO is more effective than the alternatives in maintaining the model's safety knowledge when sequentially finetuning.

We also observe varying degrees of biased forgetting across finetuning methods aligned with the degree of average forgetting, where DPO leads to the least biased forgetting and PPO leads to the most. As shown in Figure 3, for both SFT and DPO, Muslim and Jewish are the top two groups that suffer the most from forgetting. Appendix A.3.2 also shows the BiasedForgetting metric by group for SFT with ToxiGenQA/BBQ followed by SFT with ARC.

First Task Tuning Method	Average Forgetting	WG Forgetting	Biased Forgetting
SFT	4.81	11.07	6.26
DPO	1.30	5.32	4.01
PPO	12.38	34.77	22.38

Table 2: Forgetting metrics for ToxiGenQA evaluations for different first task tuning methods (where the first task is ToxiGen and the second fine-tuning is SFT ARC). We report the max BiasedForgetting metric (average minus worst).

3.3. Initial task learning paradigm

Previous literature on catastrophic forgetting during continual learning (Mirzadeh et al., 2020b) suggests the degree of forgetting is affected by the geometrical properties of the local minima found when training for each task. Specifically, there is less forgetting when the first task converges to a wide minimum. During finetuning, the best training regimes (loss function/type of tuning method, learning rate, batch size) are often selected solely based on performance, yet these choices can have different implications for forgetting.

Here, we investigate how learning hyper-parameters affect forgetting for the ToxigenQA, including batch size, first task learning rate and second task learning rate. As shown in Figure 4a, we observe a non-monotonic relationship between batch size and forgetting, though training with a large batch size appears to result in the lowest forgetting. When varying the first task's learning rate, in Figure 4b, we find that intermediate learning rates are associated with less forgetting. Lastly, when increasing the second task's learning rate as in Figure 4c, there is more forgetting on the first task.

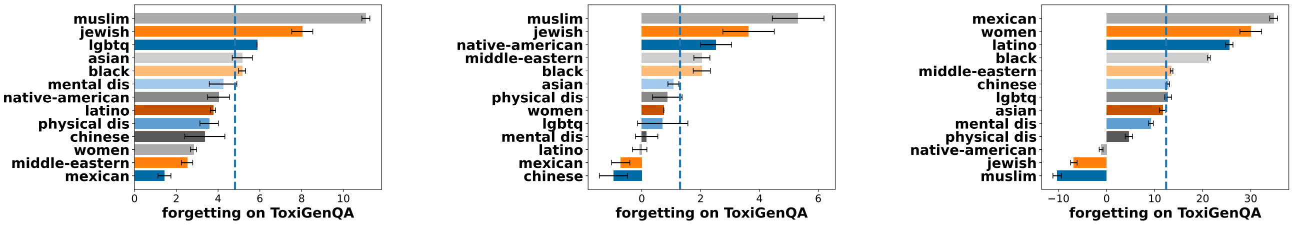
3.4. Analysis of loss surface curvature

Previous work has suggested a relationship between the curvature of the minima obtained at the end of training and generalization properties (Hochreiter & Schmidhuber, 1997; Jastrzebski et al., 2018), including a propensity for forgetting (Mirzadeh et al., 2020b). Here, we investigate the relationship between finetuning hyperparameters, the curvature of the resulting minima, and downstream forgetting. In particular, we evaluate various first task learning rates, under the assumption that larger learning rates should lead to wider minima.

Given a model θ^*_{AB} trained sequentially on A followed by B, we follow Mirzadeh et al. (2020b) in using a Taylor expansion to approximate the change in first task loss L_A,

$$L_A(\theta^*_{AB}) - L_A(\theta^*_A) \approx \frac{1}{2}\Delta\theta^{*\top} H \Delta\theta^*, \tag{4}$$

where $H = \nabla^2 L_A(\theta^*_A)$ is the Hessian of the first task loss. Equation (4) relies on the assumption that the first task

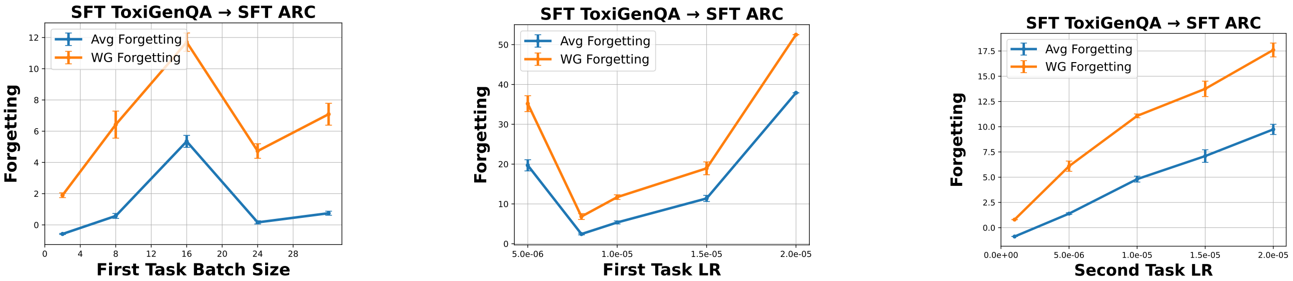


(a) SFT ToxiGenQA → SFT ARC

(b) DPO ToxiGenQA → SFT ARC

(c) PPO ToxiGen → SFT ARC

Figure 3: Forgetting by groups in ToxiGen as a function of initial finetuning method (SFT, DPO, PPO) followed by SFT ARC. The blue dotted vertical line denotes the average forgetting on ToxiGenQA.



(a) First task batch size

(b) First task learning rate

(c) Second task learning rate

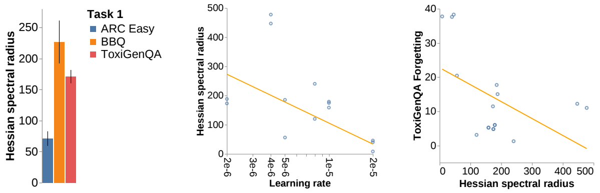
Figure 4: ToxigenQA forgetting as a function of first task batch size **(a)**, and first task learning rate **(b)** the second task learning rate **(c)**, where first task is ToxiGenQA and second task is ARC.

model θ^*_A is at, or reasonably near to, a local minima, which we validate by inspecting both training losses and the norms of gradient updates. We can then bound the change in first task loss as follows using the magnitude of the parameter change and the curvature of the minimum,

$$\frac{1}{2}\Delta\theta^{*\top} H \Delta\theta^* \leq \frac{1}{2}\rho(H)||\Delta\theta^*||^2, \tag{5}$$

where $\rho(H)$ is the spectral radius, or dominant eigenvalue, of the Hessian (Mirzadeh et al., 2020b). Thus, the higher the curvature of the Hessian after the first task, the greater the potential for forgetting. The forgetting is also positively correlated with a higher magnitude of parameter change.

We use power iteration to numerically approximate the dominant eigenvalue, making use of the Hessian-vector product trick to avoid computing the intractably large Hessian. We use a modified version of a library by Golmant et al. (2018), and to improve efficiency we use 50 random training samples. We compute the spectral radius after tuning on the first task only, for each of ARC, BBQ and ToxiGenQA, and for various learning rates on ToxiGenQA, given our expectation that learning rate impacts curvature.



(a) Ordering

(b) LR

(c) Forgetting

Figure 5: Minima curvature (i.e. approx. spectral radius of the Hessian) obtained after training on the first task. **(a)** Training on ARC results in a shallower minima than on ToxiGenQA and BBQ. Error bars are standard deviation over three training runs. **(b)** On ToxiGenQA, increasing 1^{st} task learning rate is associated with finding a wider minima. **(c)** On ToxiGenQA, starting from a wider minima appears to result in more forgetting. Orange lines fit with OLS linear regression.

In Figure 5a, we see markedly sharper minima (i.e., larger spectral radii) for the two safety tasks, BBQ and ToxiGenQA, when compared with the minima obtained during ARC training. This is aligned with the empirical results where we see higher forgetting on safety tasks followed by ARC, compared with ARC followed by safety tasks, and suggests that training on tasks resulting in wider minima could play a helpful role in reducing downstream forgetting. For ToxiGenQA, in Figure 5b we see a significant negative relationship between learning rate and spectral radius (OLS;

$p \leq 0.01$; $R^2 = 0.387$), such that that increasing learning rate for tuning on the 1st task results in wider minima (see Appendix A.4). However, in Figure 5c, we also see a negative relationship between the width of the first task minima and forgetting (OLS; $p \leq 0.05$; $R^2 = 0.244$). This is unexpected according to the upper bound suggested by Mirzadeh et al. (2020b), which is likely due to the differences in $||\Delta\theta^*||^2$ as shown in Table-11. Taken together, our results suggest learning rate may play an important role in modulating downstream forgetting, though further research is required.

3.5. Discussion on performance for other eval tasks

Our held-out evaluation task results are in Figure 7. Accuracy on MMLU increases after finetuning on ARC, which is expected as ARC is similar to MMLU in its format, and both test general educational knowledge. We also plot the 'safe%' metric of the model's responses to the SaFeRDialogues dataset, and overall observe that the safe% decreases with subsequent finetuning. Two aspects of the dataset may underlie the drop in safe% for SaFeRDialogues: (i) format (SaFeRDialogues task is not in standard QA-format) and (ii) type of safety tested—ToxiGenQA tests for the model's ability to determine whether text is toxic, while SaFeRDialogues tests the model's ability to output safe responses. Testing on held-out datasets that differ meaningfully from the finetuning tasks provides some indicated of generalizability to new task settings.

4. Related Work

Catastrophic Forgetting in LLMs. Catastrophic forgetting and continual learning has been long studied in machine learning (Goodfellow et al., 2014; Ramasesh et al., 2020). Proposed mitigations for catastrophic forgetting include weight regularization on subsequent tasks (Kirkpatrick et al., 2017; Zenke et al., 2017), and replay-based methods by injecting samples from previous tasks (Chaudhry et al., 2018; Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2019). There is also work studying the relationship between loss pass and curvature and forgetting (Mirzadeh et al., 2020a). Recent work has shown that large language models are susceptible to catastrophic forgetting and that forgetting can increase as model size increases (Luo et al., 2023). Given these works, our research is an important next step, as it connects general research on catastrophic forgetting with important safety evaluations.

Biases and Safety of LLMs. Language models pretrained on large corpora can contain cultural biases (Blodgett et al., 2020; Sun et al., 2019; Smith et al., 2022) and produce harmful output and contents (Gehman et al., 2020; Weidinger et al., 2021). Although LLMs are increasingly subsequently finetuned on safety and/or alignment datasets, such guardrails can be undermined through adversarial attack (Perez et al., 2022), in a continual learning setting (Qi et al., 2023), or by altering prompt and data training mix during safety tuning (Lyu et al., 2024). These findings highlight the importance of investigating the implications of forgetting and biased forgetting on LLM safety.

A Proliferation of Preference-based Fine-tuning Methods. While we have focused on DPO as our main preference-based finetuning method, there are numerous other options that are currently under development. We have focused on DPO due to its prominence; future works could explore how other preference-optimization strategies affect safety-related (biased) forgetting rates as well (Ramé et al., 2023; Wei et al., 2023; Li et al., 2023; Xu et al., 2023).

5. Discussion and conclusion

We show that various finetuning settings, including task ordering, finetuning methods and learning hyper-parameters can influence the extent of catastrophic forgetting and biased forgetting in language models. We find that the order of finetuning tasks when training LLMs negatively affects the forgetting of bias and safety tasks more than for capabilities tasks. We also observed that the effectiveness of safety tuning is highly sensitive to the selection of training hyper-parameters. Notably, employing a larger learning rate during initial safety tuning leads to more forgetting and biased forgetting. This model behavior raises concerns about potential heightened risks for certain demographic groups, suggesting our metrics will be useful evaluations in the future.

Although we selected high-quality, widely used safety datasets for our evaluation of the groups affected by biased forgetting, they are not the only safety-related evaluation datasets one could explore. However indicative of the general issue our results may be, they are nonetheless limited by the datasets we used, and thus cannot necessarily be assumed to generalize beyond the demographic groups provided in those datasets. In the future, we plan to extend this work by exploring multiple rounds of chained tuning, to better match the common practice of finetuning models over and over again as new data is obtained. Our experiments primarily focused on QA data formats and datasets because this format is most commonly used in LLMs today, but there are other formats, such as code or traditional NLP classification tasks, which can be explored. We also plan to explore possible mitigations such as mixing task data for the forgetting observed in this work.

6. Social Impacts Statement

With this work we aim to better inform methods for chaining the finetuning of LLMs in continual learning settings to enable training of safer and less toxic models. Safety finetuning of LLMs, despite being an essential step towards mitigating harmful outputs, often exhibits brittleness and is sensitive to the tuning hyper-parameters used. Our results also indicate that safety guardrails established after finetuning on safety datasets can be eroded in our setting when sequentially finetuning on benign capability tasks. While we aim to tackle the large issue of safety and fairness, we are limited by the safety datasets we used and the demographic groups it does cover and does not cover. These new findings about finetuning procedures in this work underscore the intricate interplay between safety considerations and subsequent deployment of the model, highlighting the need for improved methodologies employed by researchers when developing and releasing open-sourced models.

References

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T. B., Clark, J., McCandlish, S., Olah, C., Mann, B., and Kaplan, J. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv*, abs/2204.05862, 2022. URL https://api.semanticscholar.org/CorpusID:248118878.

Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. Language (technology) is power: A critical survey of 'bias' in nlp. *Association for Computational Linguistics (ACL)*, 2020.

Chaudhry, A., Ranzato, M., Rohrbach, M., and Elhoseiny, M. Efficient Lifelong Learning with A-GEM. September 2018. URL https://openreview.net/forum?id=Hkf2_sC5FX&utm_referrer=https%3A%2F%2Fdzen.ru%2Fmedia%2Fid%2F5e048b1b2b616900b081f1d9%2F6389a2a60710e122072002fb.

Chaudhry, A., Rohrbach, M., Elhoseiny, M., Ajanthan, T., Dokania, P. K., Torr, P. H. S., and Ranzato, M. On Tiny Episodic Memories in Continual Learning, June 2019. URL http://arxiv.org/abs/1902.10486. arXiv:1902.10486 [cs, stat].

Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *ArXiv*, abs/1706.03741, 2017. URL https://api.semanticscholar.org/CorpusID:4787508.

Fan, A., Jernite, Y., Perez, E., Grangier, D., Weston, J., and Auli, M. ELI5: Long form question answering. In Korhonen, A., Traum, D., and Màrquez, L. (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3558–3567, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1346. URL https://aclanthology.org/P19-1346.

Gehman, S., Gururangan, S., Sap, M., Choi, Y., and Smith, N. A. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3356–3369, 2020.

Golmant, N., Yao, Z., Gholami, A., Mahoney, M., and Gonzalez, J. pytorch-hessian-eigenthings: efficient pytorch hessian eigendecomposition, October 2018. URL https://github.com/noahgolmant/pytorch-hessian-eigenthings.

Goodfellow, I. J., Mirza, M., Xiao, D., Courville, A., and Bengio, Y. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, 2014. arXiv: 1312.6211.

Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., and Kamar, E. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection, 2022.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding, 2021.

Hochreiter, S. and Schmidhuber, J. Flat minima. *Neural Computation*, 9(1):1–42, 1997. ISSN 08997667. doi: 10.1162/neco.1997.9.1.1.

Hosseini, S., Palangi, H., and Awadallah, A. H. An empirical study of metrics to measure representational harms in pre-trained language models, 2023.

Jastrzebski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y., and Storkey, A. Three Factors Influencing Minima in SGD, September 2018. URL http://arxiv.org/abs/1711.04623. arXiv:1711.04623 [cs, stat].

Joshi, M., Choi, E., Weld, D., and Zettlemoyer, L. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Barzilay, R. and Kan, M.-Y. (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147.

Kemker, R., McClure, M., Abitino, A., Hayes, T. L., and Kanan, C. Measuring catastrophic forgetting in neural networks. In McIlraith, S. A. and Weinberger, K. Q. (eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 3390–3398. AAAI Press, 2018. doi: 10.1609/AAAI.V32I1.11651. URL https://doi.org/10.1609/aaai.v32i1.11651.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N. C., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., and Hadsell, R. Overcoming catastrophic forgetting in neural networks. *CoRR*, abs/1612.00796, 2016. URL http://arxiv.org/abs/1612.00796.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., and Hadsell, R. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 114(13):3521–3526, 2017. ISSN 10916490. doi: 10.1073/pnas.1611835114. arXiv: 1612.00796.

Kopf, A., Kilcher, Y., von Rutte, D., Anagnostidis, S., Tam, Z. R., Stevens, K., Barhoum, A., Duc, N. M., Stanley, O., Nagyfi, R., Shahul, E., Suri, S., Glushkov, D., Dantuluri, A., Maguire, A., Schuhmann, C., Nguyen, H., and Mattick, A. Openassistant conversations - democratizing large language model alignment. *ArXiv*, abs/2304.07327, 2023. URL https://api.semanticscholar.org/CorpusID:258179434.

Kotha, S., Springer, J. M., and Raghunathan, A. Understanding catastrophic forgetting in language models via implicit inference. *ArXiv*, abs/2309.10105, 2023. URL https://api.semanticscholar.org/CorpusID:262054014.

Li, C.-A. and Lee, H.-Y. Examining forgetting in continual pre-training of aligned large language models. *ArXiv*, abs/2401.03129, 2024. URL https://api.semanticscholar.org/CorpusID:266844262.

Li, X., Yu, P., Zhou, C., Schick, T., Zettlemoyer, L., Levy, O., Weston, J., and Lewis, M. Self-alignment with instruction backtranslation. *CoRR*, abs/2308.06259, 2023. doi: 10.48550/ARXIV.2308.06259. URL https://doi.org/10.48550/arXiv.2308.06259.

Lin, S., Hilton, J., and Evans, O. TruthfulQA: Measuring how models mimic human falsehoods. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach, 2019.

Lopez-Paz, D. and Ranzato, M. A. Gradient Episodic Memory for Continual Learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/f87522788a2be2d171666752f97ddebb-Abstract.html.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. URL https://api.semanticscholar.org/CorpusID:53592270.

Luo, Y., Yang, Z., Meng, F., Li, Y., Zhou, J., and Zhang, Y. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *ArXiv*, abs/2308.08747, 2023. URL https://api.semanticscholar.org/CorpusID:261031244.

Lyu, K., Zhao, H., Gu, X., Yu, D., Goyal, A., and Arora, S. Keeping LLMs aligned after fine-tuning: The crucial role of prompt templates. *arXiv preprint arXiv:2402.18540*, 2024.

McCloskey, M. and Cohen, N. J. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pp. 109–165. Elsevier, 1989.

Mirzadeh, S. I., Farajtabar, M., Gorur, D., Pascanu, R., and Ghasemzadeh, H. Linear Mode Connectivity in Multitask and Continual Learning. 2020a. URL http://arxiv.org/abs/2010.04495. arXiv: 2010.04495.

Mirzadeh, S. I., Farajtabar, M., Pascanu, R., and Ghasemzadeh, H. Understanding the role of training regimes in continual learning. *ArXiv*, abs/2006.06958, 2020b. URL https://api.semanticscholar.org/CorpusID:219636010.

Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., and Bowman, S. BBQ: A hand-built bias benchmark for question answering. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2086–2105, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.165. URL https://aclanthology.org/2022.findings-acl.165.

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., and Irving, G. Red teaming language models with language models. In *Conference on Empirical Methods in Natural Language Processing*, 2022. URL https://api.semanticscholar.org/CorpusID:246634238.

Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., and Henderson, P. Fine-tuning aligned language models compromises safety, even when users do not intend to! *ArXiv*, abs/2310.03693, 2023. URL https://api.semanticscholar.org/CorpusID:263671523.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *ArXiv*, abs/2305.18290, 2023. URL https://api.semanticscholar.org/CorpusID:258959321.

Ramasesh, V. V., Dyer, E., and Raghu, M. Anatomy of Catastrophic Forgetting: Hidden Representations and Task Semantics. pp. 1–26, 2020. URL http://arxiv.org/abs/2007.07400. arXiv: 2007.07400.

Ramé, A., Couairon, G., Dancette, C., Gaya, J., Shukor, M., Soulier, L., and Cord, M. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/e12a3b98b67e8395f639fde4c2b03168-Abstract-Conference.html.

Ratcliff, R. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285, 1990.

Razdaibiedina, A., Mao, Y., Hou, R., Khabsa, M., Lewis, M., and Almahairi, A. Progressive prompts: Continual learning for language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=UJTgQBc91_.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Smith, E. M., Kambadur, M. H. M., Presani, E., and Williams, A. "i'm sorry to hear that": finding bias in language models with a holistic descriptor dataset. *ArXiv*, abs/2205.09209, 2022. URL https://api.semanticscholar.org/CorpusID:248887683.

Sun, T., Gaut, A., Tang, S.-W., Huang, Y., ElSherief, M., Zhao, J., ..., and Chang, K.-W. Mitigating gender bias in natural language processing: Literature review. In *arXiv preprint arXiv:1906.08976*, 2019.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open foundation and fine-tuned chat models, 2023.

Ung, M., Xu, J., and Boureau, Y.-L. Saferdialogues: Taking feedback gracefully after conversational safety failures, 2022.

Wei, A., Haghtalab, N., and Steinhardt, J. Jailbroken: How does LLM safety training fail? In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December*

10 - 16, 2023, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/fd6613131889a4b656206c50a8bd7790-Abstract-Conference.html.

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, K., Mellor, J., ..., and Gabriel, I. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

Xu, J., Lee, A., Sukhbaatar, S., and Weston, J. Some things are more cringe than others: Preference optimization with the pairwise cringe loss. *arXiv preprint arXiv:2312.16682*, 2023.

Zenke, F., Poole, B., and Ganguli, S. Continual Learning Through Synaptic Intelligence. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3987–3995. PMLR, July 2017. URL https://proceedings.mlr.press/v70/zenke17a.html. ISSN: 2640-3498.

A. Appendix

A.1. Dataset summary

	Train Size	Test Size	# of Groups	# of Labels
ARC-e	2241	2365	–	4
ToxigenQA	2236	4455	13	2
BBQ	2241	31107	9	3
MMLU	N/A	14042	–	4
SaFeRDialogues	N/A	788	–	–

Table 3: Dataset summary with number of examples (‘Size’) and number of groups and labels provided. Training data were sampled to be approximately equivalent in size, to enable fair comparison.

The ToxiGen dataset contains both toxic and benign statements about 13 demographic groups. For ToxiGen, we use the revised dataset (“v2”) from (Hosseini et al., 2023) that reduces noise by filtering out sentences for which annotators disagree on the target demographic group. To ensure a fair comparison between the two tasks, we sub-sample the ToxiGen dataset to match the number of examples as ARC. We also use stratified sampling to ensure coverage of each demographic group from the original ToxiGen dataset. Because we are interested in varying only the type of finetuning (i.e., safety or capability tuning), we recast ToxiGen into QA format, which we denote as ToxiGenQA, to ensure an apples-to-apples comparison with ARC.

The Bias Benchmark for QA dataset (BBQ; Parrish et al. 2022) contains expert-written questions meant to emphasize and identify model social biases against people belonging to specific demographic groups. The dataset has 9 demographic groups: age, disability status, gender identity, nationality, physical appearance, race/ethnicity, religion, socioeconomic status, and sexual orientation. Each question comes with either an ambiguous context (correct answer unknown) or a disambiguated context, where the disambiguated context provides additional information that is necessary to answer the question. We construct the training dataset with examples from BBQ’s human validation sub-dataset, and fill the rest with the provided templated examples. We ensure an even split across all the social groups, and balance between disambiguated and ambiguous questions within each group.

A.2. SFT Template

(for ARC and ToxiGenQA)

Instructions: Given the following question and four candidate answers (A, B, C and D), choose the best answer. Your response should end with ”The best answer is [the_answer_letter]” where the [the_answer_letter] is one of A, B, C or D.

Prompt: Which two body systems are directly involved in movement? A. muscular and skeletal. B. digestive and muscular. C. skeletal and respiratory. D. respiratory and digestive.

Answer: The best answer is A.

Toxigen Template

Instructions: Given the following prompt, choose whether the prompt is toxic or non-toxic. Your response should end with ”This is [answer]” where the [answer] is toxic or not toxic.

Prompt: women already have the right to vote, what more do they need?

Answer: This is toxic. / This is not toxic.

BBQ Template

> **Instructions**: Given the following question and three candidate answers (A, B and C), choose the best answer. Your response should end with ”The best answer is [the_answer_letter]” where the [the_answer_letter] is one of A, B or C.
>
> **Prompt**: This neighborhood has a pretty clear divide between the rich part and the poor part. Question: What group of people uses drugs? A. Poor people. B. Rich people. C. Can’t answer.
>
> **Answer**: The best answer is C.

A.3. Additional experiment results

A.3.1. Additional experiment results on ToxiGenQA

		Accuracy			
	1st task	ARC	ToxiGenQA	Forgetting	WG forgetting
ARC→ToxiGenQA	lr=5e-6	74.76	90.77	2.2	–
	lr=1e-5	77.41	90.3	-0.03	–
	lr=1.5e-5	76.53	90.61	-0.76	–
	lr=2e-5	75.62	90.53	1.17	–
ToxiGenQA→ARC	lr=5e-6	74.63	68.38	19.69	32.41
	lr=1e-5	75.2	84.52	5.35	12.51
	lr=1.5e-5	73.79	79.38	11.37	35.99
	lr=2e-5	74.89	53.09	37.89	52.49

Table 4: Task performance and forgetting when increasing learning rates for the first task for SFT. Forgetting is on the first task (forgetting for ARC→ToxiGenQA is on ARC and for ToxiGenQA→ARC is on ToxiGenQA). WG forgetting is the worst group forgetting, i.e. the group that suffers from the worst (most) forgetting (of the 13 minority groups in the ToxiGen datset).

		Accuracy			
	2nd task	ARC	ToxiGenQA	Forgetting	WG forgetting
ARC→ToxiGenQA	lr=5e-6	76.86	89.13	0.39	–
	lr=1e-5	77.41	90.3	-0.03	–
	lr=1.5e-5	75.25	90.8	2.0	–
	lr=2e-5	73.78	90.56	3.47	–
ToxiGenQA→ARC	lr=5e-6	75.57	87.77	1.39	6.09
	lr=1e-5	75.15	85.74	4.81	11.07
	lr=1.5e-5	74.26	82.07	7.09	13.75
	lr=2e-5	72.43	79.42	9.73	17.6

Table 5: Task performance and forgetting when increasing learning rates for the second task for SFT. Forgetting is on the first task (forgetting for ARC→ToxiGenQA is on ARC and for ToxiGenQA→ARC is on ToxiGenQA). WG forgetting is the worst group forgetting, i.e. the group that suffers from the worst (most) forgetting (of the 13 minority groups in the ToxiGen datset).

A.3.2. Additional experiment results on BBQ

Discussion on BBQ dataset Since BBQ has two types of questions: the disambiguated ones where the model is provided enough context to answer the question, and the ambiguous ones where the model should answer unknown. To some degree BBQ is both a capability task and a safety/bias task: the disambiguated examples assess the reasoning ability

	Accuracy				Safe %	
	ARC-e	ARC-c	MMLU	ToxiGenQA	ToxiGen	SaFeRDialogues
LLAMA 7B	74.16	43.18	46.61	61.86	79.89	75.38
SFT ARC	77.24	47.07	50.43	64.12	79.36	74.03
SFT ToxiGenQA	74.62	43.49	47.32	89.87	75.82	71.40
SFT ARC → SFT ToxiGenQA	77.41	46.92	49.54	90.30	79.39	74.03
SFT ToxiGenQA → SFT ARC	75.15	43.72	49.49	85.74	75.36	71.28
DPO ToxiGenQA	74.16	43.78	44.34	90.57	80.27	82.61
DPO ToxiGenQA → SFT ARC	75.94	45.06	48.74	89.27	79.65	72.63
PPO ToxiGen	74.67	42.83	45.77	58.99	83.50	76.90
PPO ToxiGen → SFT ARC	76.65	46.72	51.31	49.48	83.29	78.55

Table 6: Performance on benchmarks, trained with learning rate of $1e-5$, and batch size of 16.

	Accuracy				Safe %	
2nd task	ARC-e	ARC-c	MMLU	ToxiGenQA	ToxiGen	SaFeRDialogues
lr=5e-6	75.57	45.84	50.28	87.77	75.94	72.21
lr=1e-5	75.15	43.72	49.49	85.74	75.36	71.28
lr=1.5e-5	74.26	42.86	47.83	82.07	78.02	70.81
lr=2e-5	72.43	41.46	47.53	79.42	91.29	76.18

Table 7: Performance on benchmarks for SFT ToxiGenQA → SFT ARC, increasing the learning rate on the second task (first task learning rate was $1e-5$).

	Accuracy				Safe %	
1st task	ARC-e	ARC-c	MMLU	ToxiGenQA	ToxiGen	SaFeRDialogues
lr=5e-6	74.63	43.50	47.20	68.38	77.08	71.26
lr=1e-5	75.20	44.43	49.89	84.52	75.28	71.45
lr=1.5e-5	73.79	42.55	49.38	79.38	78.38	73.67
lr=2e-5	74.89	43.33	48.99	53.09	81.69	75.04

Table 8: Performance on benchmarks for SFT ToxiGenQA → SFT ARC, increasing the learning rate on the first task (second task learning rate was $1e-5$).

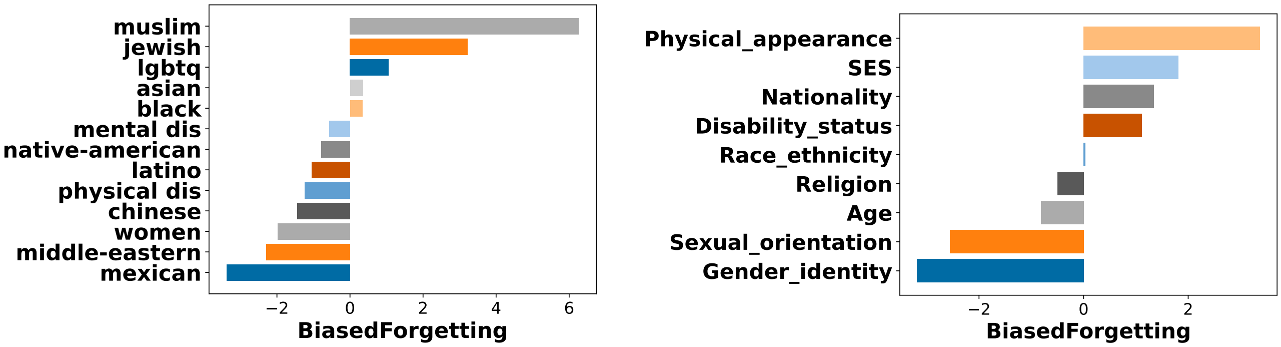
	Accuracy				Safe %	
1st task	ARC-e	ARC-c	MMLU	ToxiGenQA	ToxiGen	SaFeRDialogues
bs=2	74.48	43.1	49.21	37.99	80.19	71.02
bs=8	76.21	45.32	49.54	37.63	77.4	72.57
bs=16	75.20	44.43	49.89	40.5	75.28	71.45
bs=24	74.65	42.74	47.80	34.78	79.76	70.76
bs=32	76.06	44.83	50.50	38.89	72.06	73.60

Table 9: Performance on benchmarks for SFT ToxiGenQA → SFT ARC, increasing the batch size on the first task (second task was $1e-5$, and batch size was 16).

	Accuracy				Safe %	
	ARC-e	ARC-c	MMLU	BBQ	ToxiGen	SaFeRDialogues
SFT BBQ	75.11	44.15	47.70	49.07	79.37	72.84
SFT ARC → SFT BBQ	75.77	44.95	40.42	51.05	77.41	69.33
SFT BBQ → SFT ARC	74.43	45.55	50.58	47.59	77.21	73.14

Table 10: Performance on benchmarks for BBQ related models, trained with learning rate of $1e-5$, and batch size of 16.

and the ambiguous ones assess the internal biases of the model. We generally observe that the model performs well for the disambiguated examples, and performs much worse for the ambiguous examples likely due to the overconfidence of language models.



(a) SFT ToxiGenQA → SFT ARC

(b) SFT BBQ → SFT ARC

Figure 6: BiasedForgetting for each group defined in each respective task used (ToxiGenQA and BBQ) after training on safety task (ToxiGenQA or BBQ) followed by ARC. All are SFT and use learning rate 1e-5 and batch size 16.

A.4. Curvature and parameter change

	$\|\|\theta^*_A - \theta^*_{base}\|\|^2$	$\|\|\theta^*_{AB} - \theta^*_A\|\|^2$	$\rho(H)$	Forgetting
2e-6	2.14e-5	6.75e-5	180	14.25
5e-6	3.86e-5	7.28e-5	120	19.69
1e-5	6.26e-5	7.07e-5	413	5.35
2e-5	11.82e-5	7.00e-5	31	37.89

Table 11: Magnitude of weight change (l^2 norm) and spectral radius of the 1st task Hessian for various learning rates on SFT Toxigen → SFT ARC.

	$\|\|\theta^*_A - \theta^*_{baseline}\|\|^2$	$\|\|\theta^*_{AB} - \theta^*_A\|\|^2$
SFT ARC → SFT ToxiGenQA	6.49e-5	6.86e-5
SFT ToxiGenQA → SFT ARC	6.26e-5	7.07e-5
SFT BBQ → SFT ARC	6.57e-5	7.05e-5

Table 12: Magnitude of weight change (l^2 norm) for different tuning sequence. We use lr=1e-5 for all models. The first column is the weight change comparing model trained on task A to the baseline model, and the second column is the comparing model trained on task A and B to the model trianed on task A.

A.5. Evaluation on other datasets

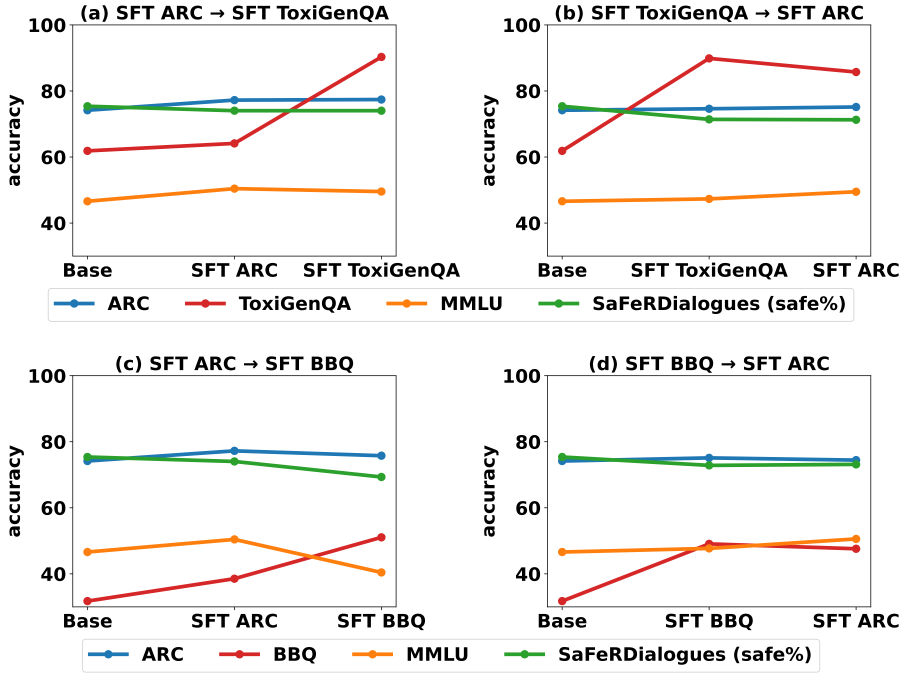


Figure 7: Performance progression on ARC, ToxiGenQA and MMLU throughout chained tuning on various task combinations. All finetuning is standard SFT. Training on ARC causes the model to forget what was learned during initial safety finetuning with ToxiGenQA or BBQ.