

MECHANISTIC UNLEARNING: ROBUST KNOWLEDGE UNLEARNING AND EDITING VIA MECHANISTIC LOCALIZATION

Phillip Guo^{1*} Aaquib Syed^{1*} Abhay Sheshadri² Aidan Ewart³ Gintare Karolina Dziugaite⁴

¹University of Maryland ²Georgia Institute of Technology ³University of Bristol

⁴Google DeepMind

{phguo, asyed04}@umd.edu

gkdz@google.com

ABSTRACT

Methods for knowledge editing and unlearning in large language models seek to edit or remove undesirable knowledge or capabilities without compromising general language modeling performance. This work investigates how mechanistic interpretability—which, in part, aims to identify model components (circuits) associated to specific interpretable mechanisms that make up a model capability—can improve the precision and effectiveness of editing and unlearning. We find a stark difference in unlearning and edit robustness when training components localized by different methods. We highlight an important distinction between methods that localize components based primarily on preserving outputs, and those finding high level mechanisms with predictable intermediate states. In particular, localizing edits/unlearning to components associated with the *lookup-table mechanism* for factual recall 1) leads to more robust edits/unlearning across different input/output formats, and 2) resists attempts to relearn the unwanted information, while also reducing unintended side effects compared to baselines, on both a sports facts dataset and the CounterFact dataset across multiple models. We also find that certain localized edits disrupt the latent knowledge in the model more than any other baselines, making unlearning more robust to various attacks.

1 INTRODUCTION

Large language models (LLMs) often learn to encode undesirable knowledge. The possibility of selectively editing or unlearning this type of knowledge is viewed as paramount for ensuring accuracy, fairness, and control of AI. Yet, editing and unlearning of knowledge from these models remains challenging.

Common editing and unlearning methods often come at the cost of affecting other general or tangential knowledge or capabilities within the model. Moreover, the edits achieved through these methods may not be robust – e.g., slight variations in the prompt formulation can often still elicit the original fact or capability, or the original answers are still present/extractable given white-box access.

Some recent work has explored editing or unlearning techniques that rely on mechanistic interpretability methods attempting to trace which components of a network store specific facts (Meng et al., 2023). These methods, such as causal tracing or attribution patching, focus on measuring how output or task accuracy is affected when clean/corrupted input is patched into specific components.

However, the effectiveness of these “output-tracing” (OT) techniques for editing has been questioned by Hase et al. (2023). Our research confirms these doubts, finding that localized editing and unlearning of facts based on several existing OT methods often perform equal to or worse than simply updating the entire model. This is particularly evident when evaluating the robustness of edits against prompt variations and relearning, and when probing for remaining latent knowledge.

*Equal contribution, determined by coin flip.

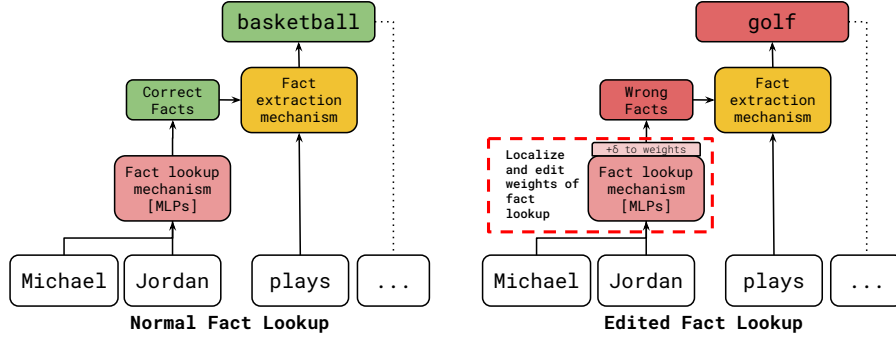


Figure 1: High level depiction of *mechanistic unlearning*. We localize components responsible for fact extraction/enrichment and modify their weights to change the associations, in order to target internal latent representations rather than targeting the output. Graph inspired by Nanda et al. (2023).

Another style of interpretability techniques first breaks down computations into high-level mechanisms with predictable intermediate states. Based on such work by Nanda et al. (2023); Geva et al. (2023), we link certain MLP layers to a fact lookup (FLU) mechanism for facts used in our analysis, that enrich the latent stream with subject attributes but don’t directly write to the output. For unlearning and edits of these facts, we only modify components that implement the FLU mechanism. More broadly, we refer to editing and unlearning that acts on components of the model identified by mechanistic intermediate component analysis as *mechanistic unlearning*. We demonstrate that FLU *mechanistic unlearning* leads to better trade-offs between edits/unlearning and maintaining performance on general language modelling capabilities, compared to edits done using OT or without any localization. Further, it exhibits improved robustness to re-learning and alternative prompting, and we demonstrate that the latent knowledge is also perturbed.

Summary of Contributions

- We perform a rigorous evaluation of several standard unlearning approaches on factual recall tasks and show they fail to generalize to prompting/output distribution shifts and adversarial relearning.
- We identify mechanisms for factual lookup and attribute extraction on Gemma-7B and Gemma-2-9B. We demonstrate that gradient-based unlearning and editing localized on the factual lookup mechanism is more robust and generalizes better than OT localizations and baselines across multiple datasets and models.
- We analyze intermediate representations using probing, and provide further evidence that editing with FLU localization reduces the available latent information more than other localizations and baselines. We also analyze the weights that are modified for each localization, and find that OT techniques and baselines modify the attribute extraction mechanisms more than the fact lookup mechanism.
- We show that editing and unlearning localized on these mechanisms is more parameter efficient, by controlling for the sizes of edits made to the model with weight masking.

1.1 RELATED WORK

Mechanistic Interpretability is a subfield of AI interpretability, aiming to understand the internal processes of AI models by attributing them to subnetworks (called circuits) within the model (Olah et al., 2020). We focus on the factual recall interpretability literature (Nanda et al., 2023; Geva et al., 2023; Chughtai et al., 2024; Yu et al., 2023), which studies methods that aim to discover mechanisms for the retrieval and formatted extraction of factual information.

Output tracing methods aim to automatically find causally important subnetworks of components for a task. Causal Tracing (Meng et al., 2023) and Automated circuit discovery (ACDC) (Conmy

et al., 2023) utilize repeated activation patching to attempt to find the subnetworks that are most critical for the model’s output on that task. Efficient methods such as attribution patching (Nanda, 2023) and edge attribution patching (Syed et al., 2023) are linear approximations of activation patching for discovering important components quickly.

Fact Editing and Machine Unlearning seek to modify pre-trained models to eliminate or alter learned knowledge such as capabilities or facts. Some prior approaches focus on identifying and removing specific individual training data points, aiming to obtain a model that is “similar” to one that had never trained on these data points (Cao & Yang, 2015; Xu et al., 2023). One formalization of unlearning to match a retrained-from-scratch model is due to Ginart et al. (2019), and is closely inspired by differential privacy (Dwork et al., 2014).

A growing body of work aims to unlearn a subset of the training data in LLMs. Eldan & Russinovich (2023) propose a method for unlearning entire books like the Harry Potter series. Chen & Yang (2023) consider modifying transformer architecture by inserting “unlearning” layers.

Fact editing focuses on overwriting factual information while preserving overall language generation ability. Meng et al. (2023) attempts to identify MLP modules that are most responsible for factual predictions via Causal Tracing and then applies a rank-one transformation upon these modules to replace factual associations.

In the context of LLMs and safety, techniques such as Helpful-Harmless RLHF (Bai et al., 2022) and Representation Misdirection for Unlearning (Li et al., 2024) aim to suppress dangerous knowledge or harmful tendencies in LLMs. A related line of work on safety proposes methods making it difficult to modify open models for use on harmful domains (Deng et al., 2024; Henderson et al., 2023).

Failures of Unlearning and Editing have been shown for both localized and nonlocalized methods. Patil et al. (2023) extract correct answers to edited facts from the intermediate residual stream and through prompt rephrasing. Yong et al. (2024) show that low-resource languages jailbreak models output unsafe content, and Lo et al. (2024); Lermen et al. (2023) demonstrate that relearning with a small amount of compute/data causes models to regain undesirable knowledge/tendencies. Even without explicit finetuning, Xhonneux et al. (2024) show that in-context learning alone suffices to reintroduce undesirable knowledge despite the model being designed to refuse to output such knowledge. Lee et al. (2024) shows that even after alignment techniques are applied to make models nontoxic, toxicity representations are still present, just not triggered - they argue that this is a reason that models lack robustness and can still be jailbroken to trigger this unwanted behavior.

2 METHODS

Our experiments are designed to test the effectiveness of localization for unlearning/editing of facts. In this section we describe the tasks used and the localization and unlearning methods evaluated.

2.1 UNLEARNING/EDITING TASKS

We focus on unlearning and editing subsets of two datasets: (1) Sports Facts dataset from Nanda et al. (2023), which contains subject-sport relations across three sports categories for 1567 athletes, and (2) the CounterFact dataset from Meng et al. (2023).

Sports Facts In the Sports Facts dataset, we attempt to unlearn and edit two groups of factual associations. First, we unlearn all athlete-sport associations for a given sport. In this case, we establish a *forget set* consisting of all the basketball athletes. Second, we edit factual associations for a set of 16 athletes belonging to all three sports categories. We test editing these sets of associations by replacing their sports with golf, and the retain set is the rest of the non-forget athletes. We use the Gemma-7B LLM (Team et al., 2024) rather than the Pythia-2.8B (Mallen & Belrose, 2023) model tested in Nanda et al. (2023), for its stronger general capabilities which we can measure for side effects, and for its ability to provide sports knowledge in different input/output formats.

CounterFact In the CounterFact dataset, following Geva et al. (2023), we first filter the dataset for facts which the model assigns higher than 50% probability to the right answer, leaving 2170

facts. Then, we edit a set of 16 facts, replacing the correct answers with an alternative false target, with the retain set being the rest of the non-forget facts. We use the Gemma-2-9B LLM, as an alternative to the Gemma-7B LLM, to test a range of model architectures and sizes. We don’t test smaller models as they cannot generally output knowledge in different formats, as measured by our robustness evaluations.

2.2 LOCALIZATION METHODS AND BASELINES

Given a model $M : X \mapsto L$ mapping sequence of tokens X to logits $L \in \mathbb{R}^{|V|}$ over vocabulary V , we consider M to be a directed acyclic graph (C, E) with C being a set of model components and E being edges between components. Adopting notation from Elhage et al. (2021), we consider the query, key, value, and output weights $W_Q^h, W_K^h, W_V^h, W_O^h$ of each head along with the input and output projection weights W_I^m, W_O^m of each MLP as components.

We are interested in finding $S : C \rightarrow \mathbb{R}$, a mapping of components to their importance in a given task. A localization is a set of components $C_\tau := \{c : c \in C, |S(c)| > \tau\}$, where τ is a threshold. In practice, we fix τ such that C_τ contains the same number of parameters in OT, FLU, and random localizations. We use these efficient localization methods for finding these mappings:

Output Tracing (OT) localization: Causal Tracing and Attribution Patching First, we test Causal Tracing, a method for finding components with high direct causal importance for factual associations (Meng et al., 2023). Previous work has highlighted the shortcomings of Causal Tracing as a localization method (Hase et al., 2023), so we also use Attribution Patching (Nanda, 2023), which uses a linear approximation to activation patching to automatically localize over components with high direct and indirect importance.

We hypothesize that these output-based techniques will prioritize the shared extraction components and other mechanisms for reformatting predictions over the more diffuse FLU components, and thus appear more precise yet leave the underlying latent information present in the model. This might decrease robustness under alternative extraction methods, thus motivating non-OT-based localization, described next.

Fact Lookup (FLU) localization: Manual Mechanistic Interpretability Next, we use manually derived localizations for MLP layers. For Sports Facts, our localization is inspired by Nanda et al. (2023), who discovered components in Pythia 2.8B responsible for *token concatenation*, *fact lookup*, and *attribute extraction*. They, along with Geva et al. (2023), find that the fact-lookup stage enriches the latent stream with information about the subject (athlete) at the subject’s token position, and the attribute extraction stage extracts the latent sport information and formats it in the final token position. We replicate a key result of their work in Gemma-7B, localizing the *fact lookup* stage to be the MLP components between layers 2 and 7.

For CounterFact, we replicate findings from Geva et al. (2023) and do further causal analyses to identify particular MLPs that perform *fact lookup* of our forget set in Gemma-2-9B. Our analysis highlights layers 3-5, 7-10, and 14-17 as the critical MLPs. The manual analysis for both datasets is outlined in Appendix A.2.1. We also locate causally important attribute extraction mechanisms, both attention heads and MLPs, in later layers of the model.

We refer to these FLU localizations as *manual interpretability*. Importantly, our analysis differs from OT techniques because we consider the causal effects of ablations upon intermediate representations used by the factual recall mechanism, not just the effects on the output. We hypothesize that the optimal location for robust unlearning/editing is in the fact lookup stage rather than in the attribute extraction stage, because adversaries can develop alternative methods for extracting knowledge from the latent stream through alternative prompts or white-box methods so we want to prevent the knowledge from ever being added to the latent stream. Thus, we exclusively modify the fact lookup MLPs.

Baselines: Random, All-MLPs, and Nonlocalized We additionally consider three baselines: one corresponding to $C_\tau = C$ (i.e., no localization, optimizing all the components of the model), another that randomly chooses components, and another that trains all MLP components. We test the last All-MLPs localization to determine if our mechanistically localized MLPs are uniquely important

- we want to know if the same unlearning performance can be achieved with just the heuristic that training only MLPs improves robustness, or if mechanistic understanding of the role of the component is crucial.

In Appendix A.2.3, we analyze the proportions of each mechanism (the extraction heads, extraction MLPs, and fact lookup MLPs, by parameter count) that are present in each localization.

2.3 PARAMETER UPDATE METHODS

Once we have a localization C_τ , we run one of the unlearning or editing methods, restricting weight updates to only components in C_τ . We modify weights using gradient ascent/descent on a combination of loss functions. As an alternative method, we also test masking a fraction of the weights as an alternative method showing precision of edits, with results reported in Section 3.2.

Localized Fine-Tuning Following work by Lee et al. (2023) and Panigrahi et al. (2023), we fine-tune the parameters within the localized components. We use a loss function

$$L = \lambda_1 L_{\text{forget}} + \lambda_2 L_{\text{retain}} + \lambda_3 L_{\text{SFT}},$$

where L_{forget} is an unlearning loss on the D_{forget} subset of facts we want to forget, L_{retain} is a cross-entropy loss on the remaining facts, and L_{SFT} is a cross-entropy loss on the Pile dataset (Gao et al., 2020). The unlearning loss we use is the $\log(1 - p)$ measure from Mazeika et al. (2024), for its stability and fewer side effects.

For editing, we use a loss function $L = \lambda_1 L_{\text{injection}} + \lambda_2 L_{\text{retain}} + \lambda_3 L_{\text{SFT}}$, where $L_{\text{injection}}$ is a cross-entropy loss on the forget facts maximizing the probability of the alternative false target. Our λ s are in Appendix A.4.

Weight Masking For weight masking-based unlearning, we train a binary differentiable mask over individual weights of the model within the localized components, inspired by weight pruning/masking work (Bayazit et al., 2023; Panigrahi et al., 2023). In this case, no weight updates are being performed. Rather, the mask turns a subset of the weights to zero. To control the sparsity penalty, we additionally include $\lambda_4 * L_{\text{reg}}$, an L1 regularization term.

3 UNLEARNING AND EDITING RESULTS

In this section, we show the results of unlearning and editing across all of the mentioned localization techniques for localized fine-tuning and weight masking. We try three unlearning/editing goals: unlearning all athletes playing basketball (referred to as unlearning sports), editing the model’s associations for a constant set of 16 athletes across all sports (akin to the error injection setup from (Hase et al., 2023), and editing the associations for 16 facts from the CounterFact dataset Meng et al. (2023). All the techniques are assessed based on standard and adversarial evaluations.

3.1 LOCALIZED FINETUNING

3.1.1 STANDARD EVALUATION

For the sports dataset, following Nanda et al. (2023), we first evaluate the accuracy of our models to complete the prompt, “Fact: [athlete] plays the sport of”, with a one-shot example of Tiger Woods playing golf given first. Note that this is the same prompt used to train the unlearning in the first place. We refer to this accuracy as Normal Accuracy. Inspired by Patil et al. (2023) and Lynch et al. (2024), we also use an alternative input and output prompting setup to measure if our unlearning has “overfitted” to the prompt input and the output format. We instead use a multiple-choice format with the choices of football, baseball, basketball, and golf. We refer to the accuracy of the model answering with the ground truth on this prompt format as the MCQ Accuracy (for both unlearning and editing, stronger methods should decrease the MCQ accuracy).

For CounterFact, we evaluate the accuracy of our models to complete the original factual associations. We report the original robustness and side effect evaluations from the Meng et al. (2023) dataset, the Paraphrase and Neighborhood facts accuracies. Since the Paraphrased prompts have the

same answer string as the original question, we hypothesize that they may be easier for model editing to generalize to - we want to test that the knowledge is edited under any expression of the knowledge. We perform a knowledge robustness check by phrasing the associations in multiple choice format (MCQ Accuracy), with the true answer, the injected false answer, and two other question-specific LLM-generated incorrect answers. Once again, accuracy is measured with the ground truth answer, so we always want the MCQ accuracy to decrease throughout editing.

Finally, we also evaluate our models’ accuracy on MMLU (Hendrycks et al., 2021) as a proxy for the general side effects of unlearning unrelated to sports and CounterFact. Our results with localized fine-tuning are shown in Table 1 for sports, Table 2 for athletes, and Table 3 for CounterFact.

Table 1: Localized fine-tuning accuracy on standard evaluations: unlearning all basketball athletes and retaining all other facts.

LOCALIZATION	FORGET ↓	RETAIN ↑	MCQ ↓	MMLU ↑
ATTRIB. PATCHING	0.000	1.000	0.767	0.602
CAUSAL TRACING	0.201	0.998	0.849	0.611
MANUAL	0.002	0.995	0.110	0.613
RANDOM	0.952	0.980	0.822	0.612
ALL-MLPs	0.000	0.994	0.279	0.606
NONLOCALIZED	0.000	0.985	0.196	0.595

Table 2: Localized fine-tuning accuracy on standard evaluations: editing a constant 16 athlete subset, considering accuracy relative to the original ground truth answers

LOCALIZATION	FORGET ↓	RETAIN ↑	MCQ ↓	MMLU ↑
ATTRIB. PATCHING	0.447	0.998	0.895	0.612
CAUSAL TRACING	0.586	0.994	0.945	0.613
MANUAL	0.001	0.970	0.108	0.611
RANDOM	0.883	0.988	0.875	0.614
ALL-MLPs	0.001	0.965	0.166	0.574
NONLOCALIZED	0.354	0.890	0.155	0.573

Table 3: Localized fine-tuning accuracy from editing 16 facts from the CounterFact dataset, considering accuracy relative to the original answers

LOCALIZATION TYPE	FORGET ↓	MAINTAIN ↑	MCQ ↓	MMLU ↑	PARAPHRASE ↓	NEIGHBORHOOD ↑
ATTRIB. PATCHING	0.000	0.974	0.944	0.691	0.286	0.681
CAUSAL TRACING	0.000	0.973	0.481	0.692	0.032	0.736
MANUAL	0.000	0.965	0.295	0.691	0.035	0.733
RANDOM	0.000	0.979	0.651	0.690	0.159	0.757
ALL-MLPs	0.000	0.927	0.390	0.688	0.083	0.725
NONLOCALIZED	0.000	0.920	0.411	0.682	0.104	0.746

As seen from the tables, across all tasks, unlearning with manual localization achieves the highest robust multiple-choice forget accuracy and near the highest MMLU, and very competitive normal forget and retain accuracy. Only manual localization, all-MLPs, and nonlocalized approaches had generalized their unlearning to the multiple choice format, but manual localization has higher MMLU performance and higher retain accuracy than All-MLPs and nonlocalized. On the CounterFact task, only manual interpretability achieves baseline MCQ accuracy, and manual localization also achieves the second-lowest score on the Paraphrase robustness evaluation as well. As before, because the Paraphrased facts may share the same attribute extraction mechanisms to the output as the original facts, we emphasize that the MCQ evaluation is the higher signal robustness evaluation.

This indicates that OT localization methods do not robustly modify the model, and the supposedly-unlearned/edited information can be extracted through prompt and task variations.

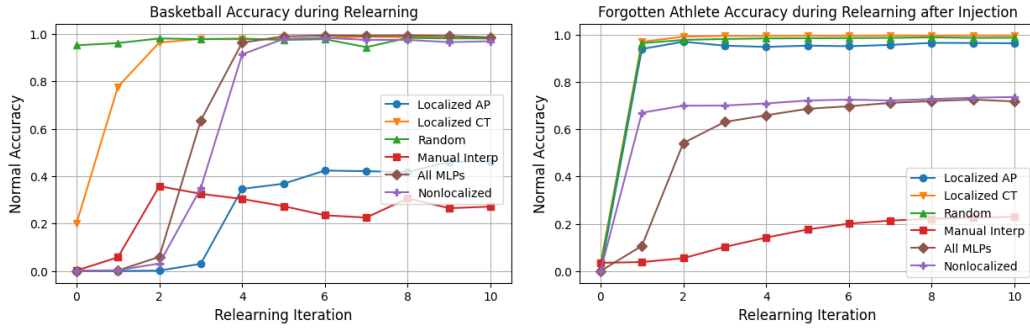


Figure 2: **(Left)** Retraining basketball-unlearned models with two athletes in the forget set, for ten iterations. **(Right)** Retraining models that have had 16 athlete-sport associations replaced with Golf as the athletes’ associated sport, with two athletes in the forget set, for ten iterations. The y-axis represents the normal accuracy on the forget set. Low-resource relearning of athletes demonstrates that using manual interpretability makes edits significantly more robust to relearning.

3.1.2 ADVERSARIAL RELEARNING

We measure the ability of our models to withstand adversarial relearning, both to address the scenario in which adversaries may have fine-tuning access and as an upper-bound measure for the quality of unlearning – a model that needs relatively fewer steps to relearn may not have properly unlearned¹. We retrain with a rank-64 LoRA across all linear modules using a subset of the forget dataset for sports, with details available in Appendix A.5.2. We measure the ability of the relearned model to answer with the ground truth statements, similar to the methodology of Deeb & Roger (2024).

Figure 2 (left) and Figure 2 (right) compare relearning robustness of different localization techniques for sports and athletes respectively.

As shown in Figure 2 (left), for sports-unlearned models, manual interpretability is the localization method that is most robust to the low-resource relearning. Unlearning based on every other localization as well as the no-localization technique regains accuracy on the rest of the forget set within a few iterations. For all of the athlete-edited models, relearning on just two of the edited athletes recovers significant accuracy on all athletes in all except for the manual interpretability model.

For CounterFact relearning, we change the setup slightly: we retrain on all forgotten statements, since the set of edited statements aren’t as directly related as for sports facts. We measure accuracy using the same train set of edited/forgotten statements in Figure 3 (left). However, because we are directly optimizing the accuracy on this train set, we also try the multiple choice evaluation on this train set to see if the relearning generalizes (if the model has reinternalized this knowledge), in Figure 3 (right). Figure 3 (right) demonstrates that the correct factual associations were not deeply relearned in any of the manual interpretability, nonlocalized, or all MLP models, with the manual interpretability model having the lowest MCQ accuracy.

3.1.3 LATENT KNOWLEDGE ANALYSIS

In this section, we provide more evidence of our hypothesis that FLU unlearning targets the source of intermediate latent knowledge rather than the extraction mechanism. We use the Sports Facts dataset for this study, as the answer is always one of three representations (the various sports).

Similar to Patil et al. (2023), we train logistic regression models (probes) (Alain & Bengio, 2018) on the activations of every model layer to predict the correct ground truth sport from the prompt,

¹Ideally, one would compare here to the number of iterations needed to learn these facts for a “gold-standard” model that was never trained on these facts during pre-training. Obtaining the gold-standard model is prohibitively expensive. Therefore, we rely on a heuristic that unlearned models that need longer re-learning behave more like the gold-standard one.

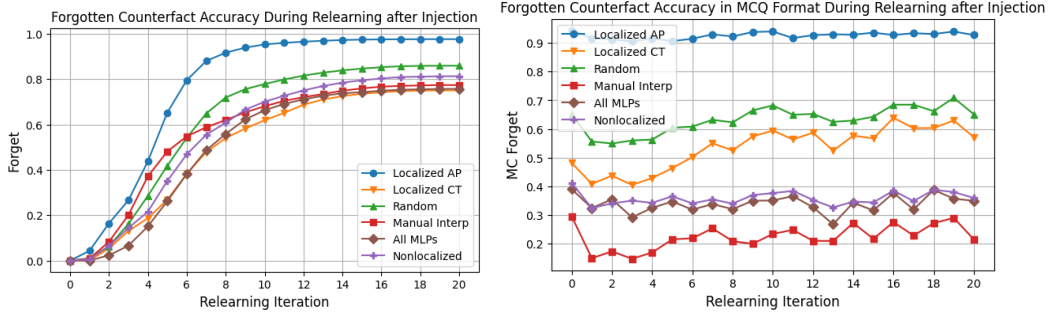


Figure 3: Retraining counterfact-edited models with all forgotten facts, for twenty iterations. **(Left)** The y-axis represents the normal accuracy on the forget set. **(Right)** The y-axis represents the Multiple Choice Formatted question accuracy on the forget set.

with the idea that a model that has truly modified a fact would not have much predictive value in its activations. For more details on probe training, see Appendix A.5.3.

We test whether the models retain information about the original athlete set in the intermediate layer representations after editing. Figure 4 shows the accuracy of the trained per-layer probes on the set of edited athletes. If model updates were concentrated in the extraction mechanism, we should still see high probe accuracies, as the probes are accessing the latent representations.

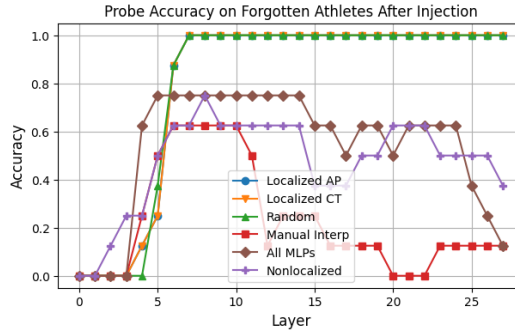


Figure 4: Probe accuracy (combined over all three sports) on the athlete golf-injected models across layers. The ‘Localized AP’, ‘Localized CT’, and ‘Random’ lines are overlapping.

Figure 4 provides evidence that the manual interpretability, All-MLPs, and nonlocalized athlete-edited models contain less recoverable representations of the edited associations in the intermediate layers, with the probes having the least accuracy on the manual interpretability model. This suggests our editing guided by manual interpretability has properly interfered with the lookup MLPs enriching the latent stream with information about the correct answer.

3.2 WEIGHT MASKING

In this section we employ weight masking to quantify the size of weight updates needed to unlearn/edit facts, for more direct comparisons. We empirically evaluate how a learned binary mask over individual weights of the localized components can produce editing/unlearning, and vary the size of this mask.

After weight masking, we also analyze the proportion of each mechanism (fact lookup, attribution extraction) that is masked for each localization in Appendix A.2.3. We demonstrate that OT methods and nonlocalized editing all modify a higher proportion of the extraction head/MLP parameters than the fact lookup mechanism parameters, supporting our claim that OT methods target extraction mechanisms rather than the fact lookup mechanisms needed for robustness.

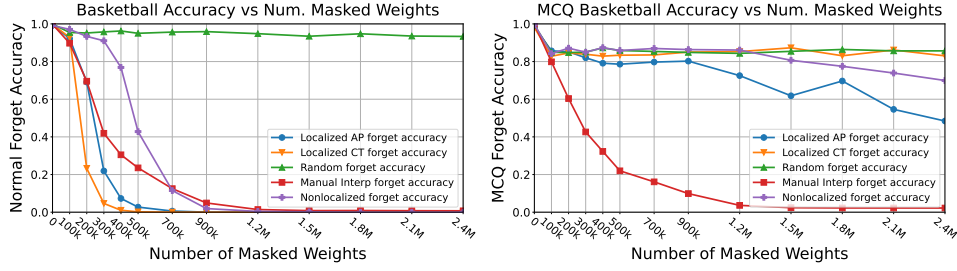


Figure 5: **(Left)** Testing the models’ unlearning of basketball athletes against the number of weights masked. **(Right)** Testing the models’ unlearning of basketball athletes against the number of weights masked, in the MCQ prompt format.

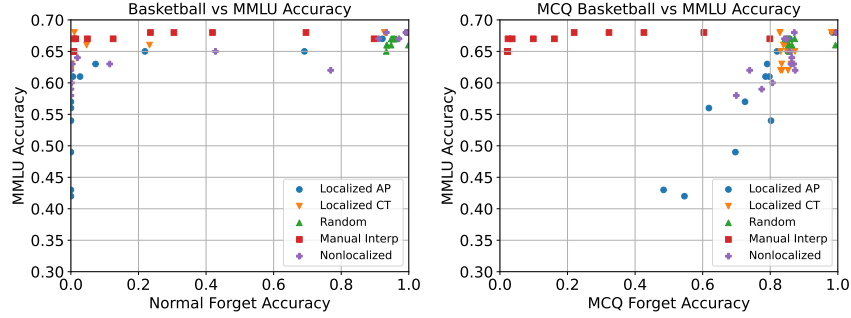


Figure 6: Unlearning basketball facts. **(Left)** Measuring MMLU and forget set performance across different discretization thresholds. **(Right)** Measuring MMLU and MCQ forget set performance across different discretization thresholds.

3.2.1 STANDARD EVALUATION

We show standard evaluations across a sweep of discretization thresholds, which directly corresponds to the size of the model edit. Figure 5 shows the accuracy on the forget and retain sets for unlearning basketball across different edit sizes. Here, we see all methods being effective in unlearning basketball facts while retaining all other facts. While AP and CT localizations cause the model to have zero accuracy on the in-distribution set with much fewer masked weights needed, when checking for generalization using a multiple-choice format we clearly see that only manual localization has successfully generalized the unlearning of basketball facts (Figure 5, right).

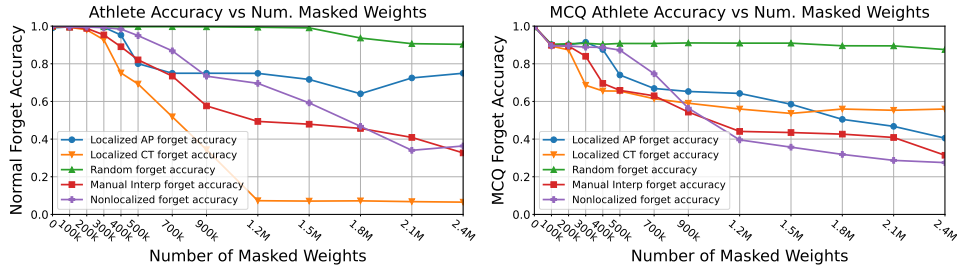


Figure 7: Editing subset of athletes. **(Left)** Measuring accuracy on the forget set. **(Right)** Measuring accuracy on the forget set in the MCQ prompt format.

We find similar results when testing for performance degradation on MMLU (because we have to evaluate many model variations, we use a smaller MMLU test set from Polo et al. (2024)). While all localized methods perform well when evaluated normally (Figure 6, left), Figure 6 (right) shows manual localization generalizes for minimizing loss of MMLU capabilities while unlearning sports facts in the MCQ format compared to the other methods.

For editing the subset of athletes, Figure 7 shows that causal tracing localization causes the model to have 0% accuracy on the forget set, and manual interpretability and nonlocalized editing cause the model to have near guessing rate (33%) accuracy. However, only manual localization minimizes loss of capabilities while editing the athlete subset (Figure 8).

Furthermore, no other method completely generalizes this unlearning to the MCQ prompt format (Figure 7), and manual localization remains superior in minimizing loss of capabilities while unlearning the athlete subset (Figure 8, right).

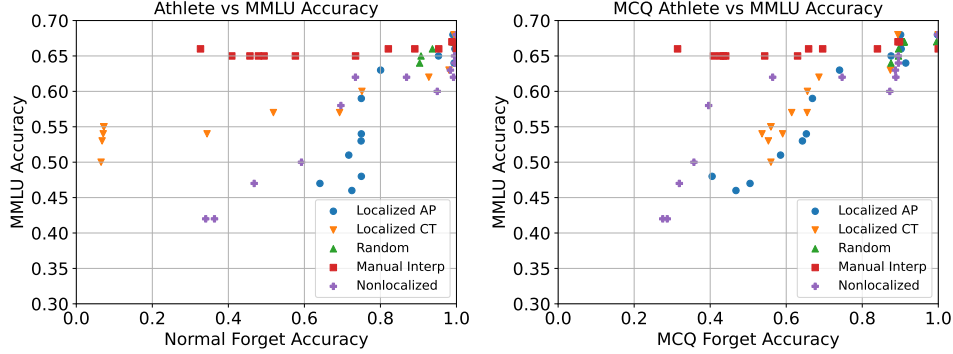


Figure 8: Editing subset of athletes. **(Left)** Measuring MMLU and forget set performance across different discretization thresholds. **(Right)** Measuring MMLU and MCQ forget set performance across different discretization thresholds.

We find similar results for editing on the CounterFact dataset. However, we find minimal difference in MMLU accuracy in all methods at all numbers of masked weights. Thus, we instead report the maintain and forget accuracies of these facts at different discretization thresholds in Figure 9. We provide additional performance metrics in Appendix A.3.

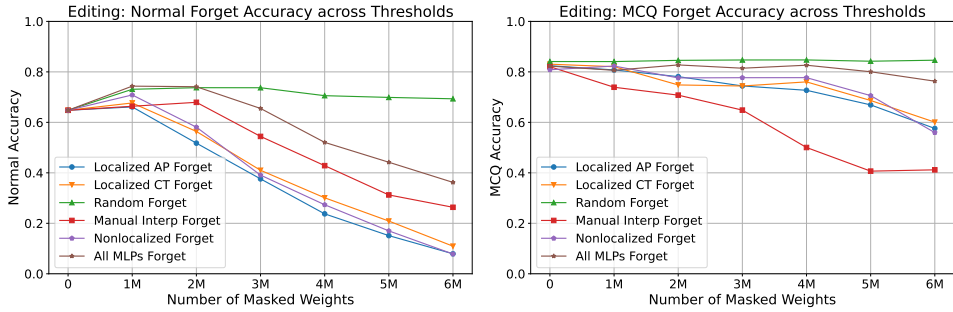


Figure 9: Editing CounterFact facts. **(Left)** Testing models accuracy on the normal forget set. **(Right)** Testing the models’ accuracy in the MCQ prompt format.

4 DISCUSSION

Recent work by Hase et al. (2023) argued that localization is not useful for model editing. Our findings demonstrate that the relationship between localization and fact editing/unlearning is more nuanced, and reveals that not all localization techniques are equal.

Our work evaluates the efficacy of different localization methods for modifying factual associations. We demonstrate clear benefits of localization for unlearning robustness through localized fine-tuning combined with manual mechanistic interpretability techniques designed for fact recall.

In Section 3.1.3 and Appendix A.2.3, we provide evidence that OT and baseline approaches fail to be robust because they target easily-localizable and high direct logit importance attention head components, that transform existing latent factual knowledge to the desired output format. This can fail

to generalize to different input and output formats and does not target the true source of knowledge in the model: other input/output formats can allow alternative attention mechanisms to transform this knowledge, and low-resource relearning can quickly repair the original attention mechanism. In contrast, FLU mechanistic understanding allows us to target unlearning at the sites where knowledge is sourced, which robustly prevents that information from entering the latent stream in any format.

Our work also suggests unlearning/editing as a potential testbed for different interpretability methods, which might sidestep the inherent lack of ground truth in interpretability (Templeton et al., 2024). We hope our work provides a framework for evaluating localizations and explanations.

ACKNOWLEDGEMENTS

We thank Stephen Casper, Aengus Lynch, and Max Li for their advice with unlearning methods and interpreting results. We also thank Daniel M. Roy, Amr Khalifa, and Eleni Triantafillou for feedback on various drafts of this work. This project used compute provided by the Center for AI Safety.

REFERENCES

- Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2018.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.
- Deniz Bayazit, Negar Foroutan, Zeming Chen, Gail Weiss, and Antoine Bosselut. Discovering knowledge-critical subnetworks in pretrained language models, 2023.
- Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pp. 463–480. IEEE, 2015.
- Jiaao Chen and Diyi Yang. Unlearn what you want to forget: Efficient unlearning for llms. *arXiv preprint arXiv:2310.20150*, 2023.
- Bilal Chughtai, Alan Cooney, and Neel Nanda. Summing up the facts: Additive mechanisms behind factual recall in llms, 2024.
- Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability, 2023.
- Aghyad Deeb and Fabien Roger. Do unlearning methods remove information from language model weights?, 2024. URL <https://arxiv.org/abs/2410.08827>.
- Jiangyi Deng, Shengyuan Pang, Yanjiao Chen, Liangming Xia, Yijie Bai, Haiqin Weng, and Wenyan Xu. Sophon: Non-fine-tunable learning to restrain task transferability for pre-trained models. *arXiv preprint arXiv:2404.12699*, 2024.
- Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.
- Ronen Eldan and Mark Russinovich. Who’s harry potter? approximate unlearning in llms. *arXiv preprint arXiv:2310.02238*, 2023.
- Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. <https://transformer-circuits.pub/2021/framework/index.html>.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling, 2020.
- Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models, 2023.
- Antonio Ginart, Melody Y. Guan, Gregory Valiant, and James Zou. Making ai forget you: Data deletion in machine learning, 2019.
- Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models, 2023.
- Peter Henderson, Eric Mitchell, Christopher Manning, Dan Jurafsky, and Chelsea Finn. Self-destructing models: Increasing the costs of harmful dual uses of foundation models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 287–296, 2023.

-
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021.
- Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.
- Andrew Lee, Xiaoyan Bai, Itamar Pres, Martin Wattenberg, Jonathan K Kummerfeld, and Rada Mihalcea. A mechanistic understanding of alignment algorithms: A case study on dpo and toxicity. *arXiv preprint arXiv:2401.01967*, 2024.
- Yoonho Lee, Annie S. Chen, Fahim Tajwar, Ananya Kumar, Huaxiu Yao, Percy Liang, and Chelsea Finn. Surgical fine-tuning improves adaptation to distribution shifts, 2023.
- Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish. Lora fine-tuning efficiently undoes safety training in llama 2-chat 70b, 2023.
- Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassan Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhruhu Bharathi, Adam Khoja, Zhenqi Zhao, Ariel Herbert-Voss, Cort B. Breuer, Samuel Marks, Oam Patel, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Russell Kaplan, Ian Steneker, David Campbell, Brad Jokubaitis, Alex Levinson, Jean Wang, William Qian, Kallol Krishna Karmakar, Steven Basart, Stephen Fitz, Mindy Levine, Ponurangam Kumaraguru, Uday Tupakula, Vijay Varadharajan, Ruoyu Wang, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The wmdp benchmark: Measuring and reducing malicious use with unlearning, 2024.
- Michelle Lo, Shay B. Cohen, and Fazl Barez. Large language models relearn removed concepts, 2024.
- Aengus Lynch, Phillip Guo, Aidan Ewart, Stephen Casper, and Dylan Hadfield-Menell. Eight methods to evaluate robust unlearning in llms, 2024.
- Alex Mallen and Nora Belrose. Eliciting latent knowledge from quirky language models, 2023.
- Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal, 2024.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt, 2023.
- Neel Nanda. Attribution patching: Activation patching at industrial scale, 2023. URL <https://www.neelnanda.io/mechanistic-interpretability/attribution-patching>.
- Neel Nanda, Senthooan Rajamanoharan, János Kramár, and Rohin Shah. Fact finding: Attempting to reverse-engineer factual recall on the neuron level, Dec 2023. URL <https://www.alignmentforum.org/posts/iGuwZTHWb6DFY3sKB/fact-finding-attempting-to-reverse-engineer-factual-recall>.
- Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001. <https://distill.pub/2020/circuits/zoom-in>.
- Abhishek Panigrahi, Nikunj Saunshi, Haoyu Zhao, and Sanjeev Arora. Task-specific skill localization in fine-tuned language models, 2023.
- Vaidehi Patil, Peter Hase, and Mohit Bansal. Can sensitive information be deleted from llms? objectives for defending against extraction attacks, 2023.
- Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples, 2024.

-
- Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery, 2023.
- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.
- Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermy, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL <https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html>.
- Sophie Xhonneux, David Dobre, Jian Tang, Gauthier Gidel, and Dhanya Sridhar. In-context learning can re-learn forbidden tasks. *arXiv preprint arXiv:2402.05723*, 2024.
- Heng Xu, Tianqing Zhu, Lefeng Zhang, Wanlei Zhou, and Philip S Yu. Machine unlearning: A survey. *ACM Computing Surveys*, 56(1):1–36, 2023.
- Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. Low-resource languages jailbreak gpt-4, 2024.
- Qinan Yu, Jack Merullo, and Ellie Pavlick. Characterizing mechanisms for factual recall in language models, 2023.

A APPENDIX

A.1 FACT INJECTION RESULTS

We additionally consider the common factual editing methodology, specifically an error injection setup (Hase et al., 2023) where we replace correct athlete-sport associations with incorrect associations between athlete and "Golf". The results are in Table 4 and Table 2, where we demonstrate that our manual localization method again achieves the strongest robust unlearning generalization while maintaining more general capabilities than the other robustly unlearned model.

Table 4: Results of unlearning basketball associations, with the objective of replacing the correct sport of "Basketball" with "Golf". Forget refers to the model’s accuracy at stating the original sport association which should have been replaced.

LOCALIZATION	FORGET ↓	RETAIN ↑	MCQ ↓	MMLU ↑
ATTRIB. PATCHING	0.000	1.000	0.815	0.611
CAUSAL TRACING	0.028	1.000	0.866	0.614
MANUAL	0.035	0.973	0.257	0.610
RANDOM	0.018	1.000	0.839	0.611
ALL-MLPs	0.000	0.946	0.363	0.571
NONLOCALIZED	0.000	0.995	0.376	0.565

Table 5: Localized fine-tuning accuracy on standard evaluations: Unlearning a constant 16 athlete subset, retaining all other facts.

LOCALIZATION	FORGET ↓	RETAIN ↑	MCQ ↓	MMLU ↑
ATTRIB. PATCHING	0.941	0.964	0.934	0.614
CAUSAL TRACING	0.891	0.915	0.910	0.612
MANUAL	0.034	0.975	0.175	0.615
RANDOM	0.938	0.952	0.883	0.612
ALL-MLPs	0.003	0.973	0.281	0.599
NONLOCALIZED	0.203	0.570	0.391	0.540

We also perform relearning and latent knowledge experiments in Figure 2 (right) and Figure 4 (Appendix A), demonstrating that manual localization for the athlete subset injection improves relearning and latent knowledge robustness.

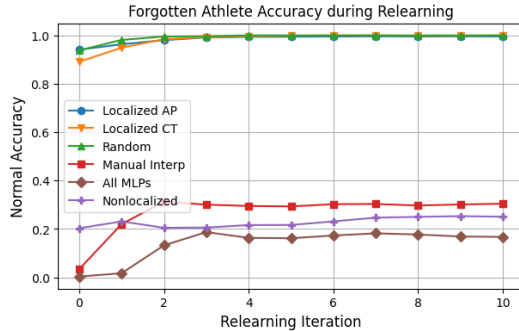


Figure 10: Retraining athlete-unlearned models with two athletes in the forget set, for ten iterations. The y-axis represents the normal accuracy on the forget set. Low-resource relearning of athletes demonstrates that manual and non-localized unlearning techniques are robust to this test (staying close to the guessing rate of 33%), while the other methods maintain full performance on the entire forget set.

A.2 GEMMA INTERPRETABILITY ANALYSIS

A.2.1 SPORTS FACTS

We redo analysis from Nanda et al. (2023) on Gemma-7B. We train logistic regression models (“probes”) to predict the correct sport given the internal representation of the model at a layer. We find that probes predicting the correct sport increase in accuracy significantly in layers 2 through 7, and we find the mean ablation of all attention heads past layer 7 to have minimal impact on the linear representation of player attributes (Figure 12).

Unlike Nanda et al. (2023), however, we find attention heads past layer 2 that impact the linear representation of attributes and thus could potentially be important for fact lookup (Figure 11). However, because they could likely play a variety of other different roles such as token concatenation, following the findings of Geva et al. (2023); Nanda et al. (2023) that MLPs do primary factual representation enrichment, in this work we only consider the MLPs as our localization.

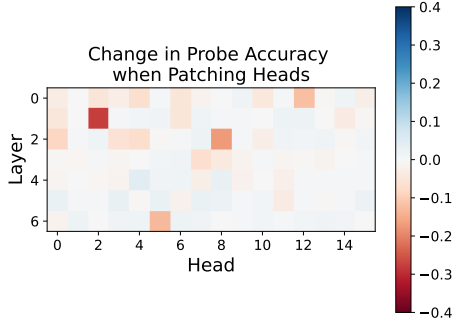


Figure 11: Difference in final layer probe accuracy when mean ablating a single head for all heads between layers 0 and 6.

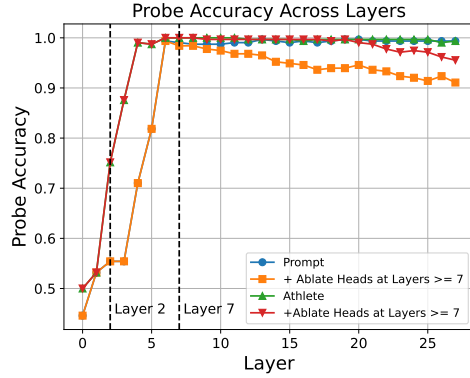


Figure 12: Probe accuracy on predicting sport across layers. “Prompt” refers to the entire facts prompt, while “Athlete” is just the athlete’s name.

A.2.2 COUNTERFACT

We repeat analysis from Nanda et al. (2023) and Geva et al. (2023) on Gemma-2-9B. We first measure the effect on the difference in logits between correct and incorrect answers of facts when patching the direct path of attention heads and MLPs to the final output, shown in Figure 13 and Figure 14. An attention head or MLP will have a large effect on the logit difference if it is important in moving the factual information to the last token position or decoding it into the correct answer. We call these components part of the “fact extraction mechanism”, and aim to find the source of the factual information moved by this mechanism.

To find this source, we patch the outputs of MLPs to this “fact extraction mechanism” and measure the resultant change in logit difference (Figure 15). An MLP would cause a large change in logit difference if it caused relevant representations to form that are then moved by the “fact extraction heads” to increase the probability of the correct output. We take the MLPs with the highest change and include them in our manual localization of CounterFact (MLPs 3-5, 7-10, 14-17).

Change in Logit Difference, Path Patching Heads -> Final Output

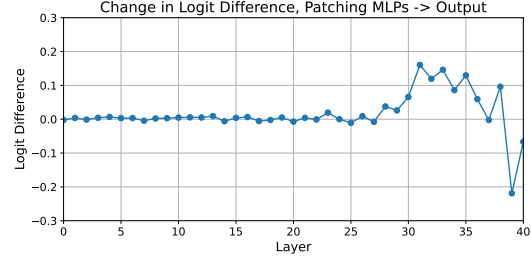
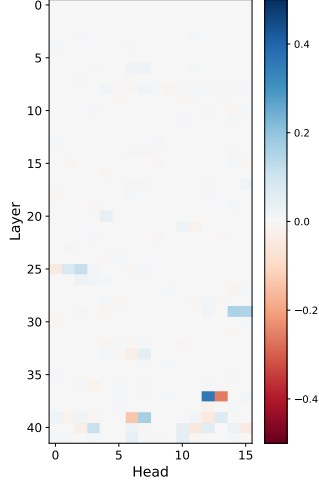


Figure 14: Change in logit difference when patching MLPs to the final output.

Figure 13: Change in logit difference when patching heads to the final output.

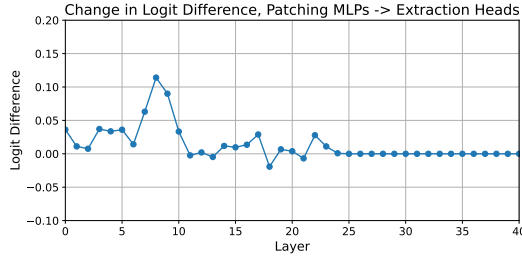


Figure 15: Change in logit difference when patching MLPs to the extraction mechanism.

A.2.3 MECHANISM WEIGHT ANALYSIS

We analyze the actual components localized by each localization type and our baselines, for the CounterFact editing task. We seek to demonstrate that the OT localizations and baselines target extraction mechanisms rather than just the FLU mechanisms.

First, in Table 6, we compare the parameter counts of the part of each mechanism that is present in each localization. Table 6 shows that causal tracing and attribution patching both have the potential to modify a considerable proportion of the extraction heads and extraction MLPs.

Then, in Table 7, we compare the proportion of each mechanism that is masked when using a localized weight mask and discretizing to about 6 million weights. This is one approximate metric for how much each mechanism is modified by the localized editing. Table 7 demonstrates that attribution patching, causal tracing, and nonlocalized editing all modify a higher proportion of the extraction head/MLP weights than the fact lookup mechanism weights.

This supports our argument that OT methods target high logit-diff extraction mechanisms, rather than the fact lookup mechanisms that enrich the latent stream with the correct attributes, which decreases the robustness of edits/unlearning. It is important to note that since our manual interpretability localization is based on our discovered mechanisms, this does not serve as an evaluation of manual interpretability, but rather only of causal tracing and attribution patching.

Table 6: Comparison of total parameters of each mechanism that are present in each localization, for editing 16 facts from CounterFact

LOCALIZATION	EXTRACTION HEADS	EXTRACTION MLPs	FACT LOOKUP
TOTAL	27,448,320	1,027,604,480	1,130,364,928
ATTRIB. PATCHING	13,724,160 (50.0%)	616,562,688 (60.0%)	102,760,448 (9.1%)
CAUSAL TRACING	8,234,496 (30.0%)	308,281,344 (30.0%)	411,041,792 (36.4%)
MANUAL INTERP.	0	0	1,130,364,928 (100.0%)
ALL-MLPs	0	1,027,604,480 (100.0%)	1,130,364,928 (100.0%)
NONLOCALIZED	27,448,320 (100.0%)	1,027,604,480 (100.0%)	1,130,364,928 (100.0%)

Table 7: Comparison of parameters of each mechanism that are masked by a trained weight mask, discretized to about 6 million weights

LOCALIZATION TYPE	EXTRACTION HEADS	EXTRACTION MLPs	FACT LOOKUP
TOTAL (BASELINE)	27,448,320 (100%)	1,027,604,480 (100%)	1,130,364,928 (100%)
ATTRIB. PATCHING	165,300 (0.60%)	1,479,877 (0.14%)	1,385,198 (0.12%)
CAUSAL TRACING	30,828 (0.11%)	1,491,040 (0.15%)	1,424,059 (0.13%)
MANUAL INTERP.	0 (0.0%)	0 (0.0%)	6,248,039 (0.55%)
ALL-MLPs	0 (0.0%)	1,378,744 (0.13%)	1,663,772 (0.15%)
NONLOCALIZED	358,918 (1.3%)	1,198,211 (0.12%)	1,174,939 (0.10%)

A.3 ADDITIONAL COUNTERFACT WEIGHT MASKING RESULTS

We report a comparison of all localizations across discretization thresholds for normal and MCQ forget sets in Figure 16 and Figure 17. We see that manual interpretability outperforms all other methods of localization in preserving maintain accuracy while decreasing forget accuracy.

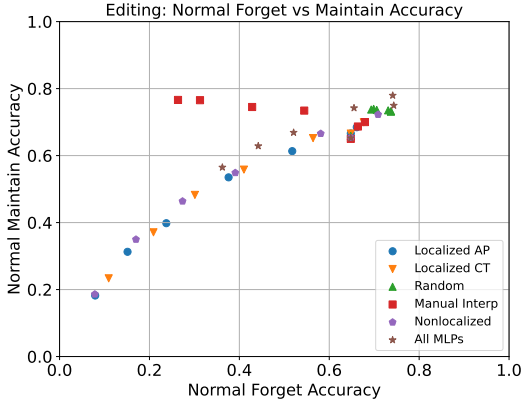


Figure 16: Accuracy on normal forget set vs on the maintain set across localizations and discretization thresholds.

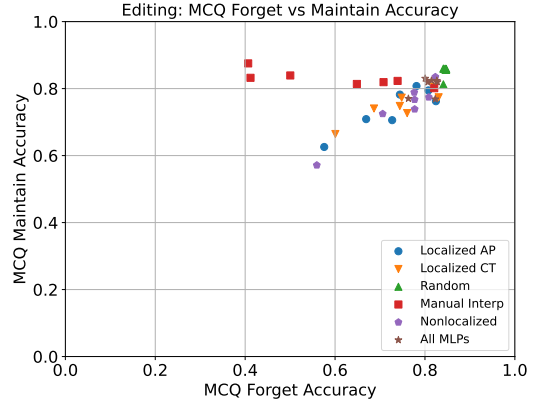


Figure 17: Accuracy on multiple choice input vs on the maintain set across localizations and discretization thresholds.

We perform additional adversarial analysis of accuracies across different discretization thresholds. We report the "paraphrase" and "neighborhood" adversarial results in Figure 18 and Figure 19.

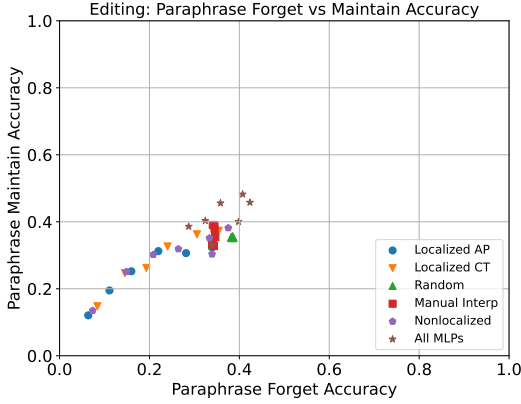


Figure 18: Accuracy on paraphrased input vs on the maintain set across localizations and discretization thresholds.

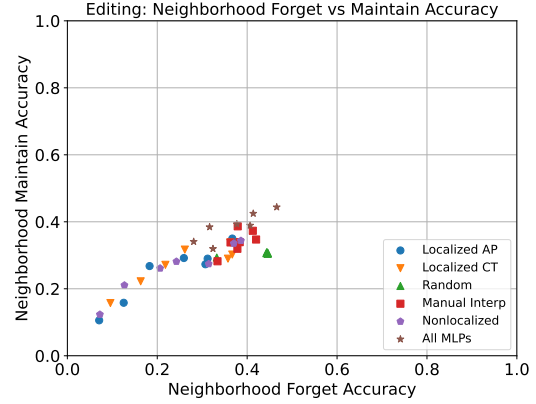


Figure 19: Accuracy on "neighborhood" input vs on the maintain set across localizations and discretization thresholds.

A.4 HYPERPARAMETERS

For localized fine-tuning, we use $\lambda_1 = 0.2$ for forgetting basketball associations and $\lambda_1 = 1$ for forgetting particular athlete associations. For weight masking on Sports Facts, we use $\lambda_1 = 0.3$ for both basketball and athlete associations. λ_2 and λ_3 , the retain and SFT loss coefficients, were both set to 1 across all experiments. For weight masking regularization, we used $\lambda_4 = 1 * 10^{-7}$ (where our regularization loss was the total sum of absolute weight mask values).

For weight masking on CounterFact, we use $\lambda_1 = \lambda_2 = \lambda_3 = 1$. For weight masking regularization, we again used $\lambda_4 = 1 * 10^{-7}$ (where our regularization loss was the total sum of absolute weight mask values).

For localized fine-tuning on Gemma, we trained using 50 iterations of batch size 4 with 16 accumulation steps, using an AdamW optimizer (Kingma & Ba, 2017) with 0 weight decay, a learning rate of $1 * 10^{-5}$, and a cosine annealing scheduler for both basketball and athletes.

For weight masking, we unlearned using 50 iterations of batch size 10 with 15 accumulation steps, using an SGD optimizer (for memory efficiency) with learning rate of $1 * 10^{-2}$ and clamping the mask values between 0 and 1 every update step.

A.5 EVALUATION DETAILS

A.5.1 TRAIN-TEST SPLITS

We split the basketball D_{forget} set and both D_{retain} sets (basketball and athletes) into an 80%-20% train-test-split, and all of our reported numbers are on the test set. We do not split the D_{forget} set of 16 athletes, because we wish to test if the model has unlearned the athletes it was trained to unlearn.

A.5.2 DETAILS ON ADVERSARIAL RELEARNING

We retrain the model with only two athletes in D_{forget} for multiple iterations (along with a standard retain and SFT loss), in both the sport and athlete unlearning scenarios. In practice, for basketball this looks like giving the model the same batch of only Boris Diaw and Jae Crowder multiple times, and for athletes we give a batch of only DeForest Buckner and Walter Payton. We retrain with a rank-64 LoRA on all linear modules.

A.5.3 DETAILS ON LATENT KNOWLEDGE

We don't follow the same methodology as Patil et al. (2023) because we only care about the same three possible tokens, so it isn't applicable to apply their search-budget methodology and we instead try linear probes. In general, we don't consider linear probing to be a realistic threat model for

beating unlearning, as attackers need white-box access and labels for large subsets of the forget set, but we do these tests for an approximate upper bound of accessible information by a capable-enough adversary.

We train three linear probes (Alain & Bengio, 2018) for every model and layer, one for each sport (to predict True or False with a base rate of 66%), on samples from both the forget and retain datasets. We train each probe on both forget and retain samples because for sports, there is only one forget sport (so the answer would be constant if we trained different probes), and for athletes we only have 16 total examples that must further be split into train-test.

For athletes, we split the forget set into a 50%-50% train-test split, so the probe training dataset includes 8 of the forgotten athletes (along with the standard retain train split) and the test set includes the other 8 (along with the retain test split). For sports, we use the standard basketball train and test split. Then, as a measure of aggregated accuracy, we only consider a test sample to be correct if probes for all three sports are correct.