
Mechanistic Unlearning: Robust Knowledge Unlearning and Editing via Mechanistic Localization

Phillip Guo^{*1} Aaqib Syed^{*2} Abhay Sheshadri³ Aidan Ewart⁴ Gintare Karolina Dziugaite⁵

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.

We coin a new term to categorize localizations which measure causal effects of components on only the output: *Output-Tracing* localizations. The effectiveness of output-tracing (OT) techniques like Causal Tracing 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.

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.

^{*}Equal contribution ¹Work done while at University of Maryland, College Park ²University of Maryland, College Park ³Georgia Institute of Technology ⁴University of Bristol ⁵Google DeepMind. Correspondence to: Aaqib Syed <asyed04@umd.edu>, Phillip Guo <phguo@umd.edu>, Gintare Karolina Dziugaite <gkdz@google.com>.

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 editing approaches on factual recall tasks, and we identify mechanisms for factual lookup and attribute extraction on Gemma-7B, Gemma-2-9B, and Llama-3-8B. We demonstrate that gradient-based editing localized on the factual lookup mechanism is more robust than OT localizations and baselines across multiple datasets, models, and evaluations.
- We demonstrate that it is more difficult to elicit the forgotten ground truth answers using alternative prompting with FLU localizations. We also demonstrate slower or no relearning of the ground truth answers, retraining edited models on half of the edited set and evaluating them on the other half of the edited set.
- We analyze intermediate representations using probing, and provide further evidence that editing with FLU localization modifies the internal latent information to reflect the desired edited answer 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).

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., 2024b) aim to suppress dangerous knowledge or harmful tendencies in LLMs. Li et al. (2024a); Zou et al. (2023; 2024) approach unlearning and dangerous knowledge suppression from a top-down feature view, reading or suppressing linear features related to memorized, harmful, and undesired concepts. A related line of work on safety proposes methods making it difficult to modify open models for use on harmful domains (Tamirisa et al., 2024; Deng et al., 2024; Henderson et al., 2023), including through adversarial relearning.

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); Deeb & Roger (2024) 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

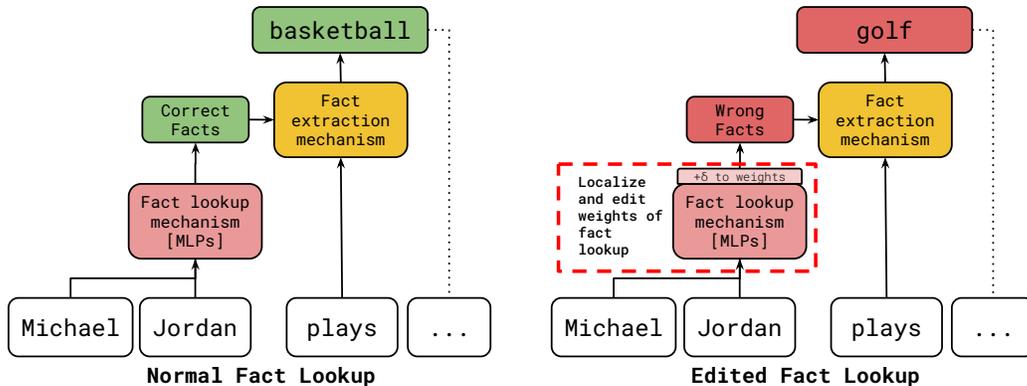


Figure 1. High level depiction of *mechanistic unlearning*. We localize components responsible for fact extraction/enrichment and modify their weights to change the associations, in order to target internal latent representations rather than targeting the output. Graph inspired by Nanda et al. (2023).

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.

Hong et al. (2024) evaluates unlearning by measuring residual knowledge left in internal activations, and demonstrate that current approaches fail to remove this residual knowledge and thus can be exploited. They attempt unlearning by targeting the MLPs these residual knowledge traces reside in, but fail to find a non-oracle unlearning approach that successfully removes residual information.

2. Methods

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

2.1. Editing Tasks

We focus on 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: Sports-Athlete-Editing, Full-Sports-Editing, and Sports-Unlearning tasks In the Sports Facts dataset, we edit two general groups of factual associations.

For the first editing task, we edit factual associations for a constant set of randomly selected athletes belonging to any of the three sport categories. We test editing these sets of associations by replacing their correct sports with one of the other two incorrect sports (with equal probability).

To increase the comprehensiveness of our evaluation, we run experiments with different *forget set* sizes: 16 athletes and 64 athletes. We refer to this task as **Sports-Athlete-Editing**. These chosen forget sets are constant between all localizations. For the second editing task, we unlearn all athlete-sport associations for athletes who play one sport. In this case, we establish a forget set consisting of all the athletes who play one sport (basketball, baseball, or football), and we edit the association by replacing the athlete’s correct sport with golf. For comprehensiveness, we vary the sport that the forget set is constructed from. We refer to this task as **Full-Sports-Editing**. Finally, we also design an unlearning task, **Sports-Unlearning**, where the goal is to unlearn associations for all athletes in one of the sports.

CounterFact-Editing and Sequential-CounterFact-Editing task In the CounterFact dataset, following Geva et al. (2023), we first filter the dataset for facts which our models assign higher than 50% probability to the right answer, which varies per model. The goal of our **CounterFact-Editing** task is to edit a constant set of facts, replacing the correct answers with an alternative false target, with the retain set being the rest of the non-forget facts. We vary forget set sizes to be of 16 and 64 facts. In **Sequential-CounterFact-Editing** task, we edit a total of 64 facts by sequentially editing four randomly selected subsets of 16 facts. We test sequential editing here because facts from CounterFact can reside in different parts of the model, so we wish to test if we can exploit different localizations for different facts. These chosen forget sets are constant between all localizations. We also scaled this approach to 1000 facts while creating localizations for each fact individually, showing a more automated approach to edit facts robustly at scale.

Models We implement editing on the Gemma-7B LLM, the Gemma-2-9B LLM, and the Llama-3-8b LLM. We don’t use the Pythia-2.8B (Biderman et al., 2023) and GPT-2 models tested in the previous fact interpretability literature because our larger models have stronger general capabilities which we can measure for side effects, and also because our larger models can provide factual knowledge in more input/output formats for more robustness evaluations. All graphs presented in the main text and appendix, unless otherwise specified, are averaged over all three model types.

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 of each head along with the input, gate, and output projection weights 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). We also use Attribution Patching (Nanda, 2023) as a fast and acceptably accurate approximation of causal tracing to automatically localize over components with high direct and indirect importance. We additionally consider the versions of these localizations with only MLPs (*Causal Tracing MLPs* and *Attribution Patching MLPs*).

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. We discuss the precise components/mechanisms highlighted by OT localizations in Appendix A.3.

Fact Lookup (FLU) localization: Next, we use manually derived localizations for MLP layers. For Sports Facts, our localization is inspired by Nanda et al. (2023), who identified components in Pythia 2.8B responsible for *token*

concatenation, *fact lookup*, and *attribute extraction*. Their work, along with Geva et al. (2023), demonstrates that the fact-lookup stage enriches the latent representation of the subject (athlete) with information about their corresponding sport, while 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 our three models by training logistic regression models (probes) to predict the correct sport using the internal activations of the model taken from different layers. We consider the FLU localization to be the layers at which the probe accuracies rapidly increase as these correspond to layers where representations of the athletes are being enriched to encode the correct sport.

For CounterFact, we cannot use the probe technique since, unlike Sports Facts, the correct answers for the dataset do not fall into a few categories that we can train a probe for. Instead, we first use path patching from Goldowsky-Dill et al. (2023) to measure the direct importance of attention heads and MLPs on the final logit difference between the correct answer and a sample incorrect answer. Path patching finds the effect of corrupting a single path from a sender component to a receiver component on the final logits of the model. Components that change the output significantly without being mediated by other components do so by directly affecting the logits, and thus we consider them to be responsible for extracting the facts encoded in the representation into the answer logits. We thus refer to these components as the “attribute extraction mechanism” Geva et al. (2023). We use path patching again, this time patching paths between MLPs and the attribute extraction mechanism to find the components with the highest contribution to the logit difference as mediated through the extraction mechanisms. Such components enrich the token representations with the appropriate facts to then be extracted, and thus we use them as our FLU localization. More details about the manual analysis for both datasets is outlined in Appendix A.2.1.

Importantly, FLU differs from OT techniques by focusing on 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 editing is in the fact lookup stage rather than in the attribute extraction stage, as adversaries could potentially devise alternative extraction methods if knowledge remains in the latent stream. Therefore, our model edits are focused exclusively on the fact lookup MLPs.

Baselines: Random-MLPs, Random, All-MLPs, and Nonlocalized We additionally consider four baselines: one corresponding to $C_\tau = C$ (i.e., no localization, optimizing all the components of the model), another that randomly chooses components, another that trains all MLP components, and another trains a random selection of MLP

components. We test the last MLP baselines 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.5, 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.

The main text focuses on comparing FLU, Causal Tracing, and Nonlocalized, while the appendix has the same figures with all above localizations included, with the same conclusions in every case.

2.3. Parameter Update Method

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 update weights using gradient descent on a combination of loss functions.

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

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 new false target. L_{retain} is a cross-entropy loss on a train split of the remaining facts, and L_{SFT} is a cross-entropy loss on the Pile dataset (Gao et al., 2020). We sweep over learning rates and injection loss λ s for three representative localizations in Appendix A.6.

3. Editing Evaluation

In this section, we show the results of model editing with localization: we test localization techniques from Section 2.2, and edit these localized components using fine-tuning (Section 2.3). Here we focus on four main editing tasks: **sports-athlete-editing**, **full-sports-editing**, **counterfact-editing**, and **sequential-counterfact-editing**. We present augmenting results for **sports-unlearning** in Appendix A.1.

All the editing tasks are assessed based on prompt-completion based and adversarial relearning evaluations.

3.1. Prompting-based Evaluation

Our prompting-based evaluation assesses an editing method’s ability to forget or edit specific information while retaining unrelated knowledge, measured by evaluating how the model post-editing completes the prompts coming from the forget set. We report how accurately it recalls the unde-

sired forgotten answer (*forget accuracy*), and how accurately it recalls the new desired edited answer (*edit accuracy*). In addition, we also measure the accuracy on facts not in the forget set (*maintain accuracy*). In cases where a positive result is lower accuracy, we use the term *error* to denote $1 - \text{accuracy}$ (e.g. *forget error* = $1 - \text{forget accuracy}$). Well-edited models should decrease forget accuracy, increase forget error, and increase edit accuracy. Results of these standard evaluations are reported in Appendix A.7.1.

Inspired by Patil et al. (2023) and Lynch et al. (2024), to ensure the editing process has not overfit to the specific format of the original prompts, we incorporate a robustness check using a multiple-choice question format (MCQ accuracy). This helps determine to what extent the model edited the information, and whether it can still access and utilize that knowledge when prompted differently. In this MCQ evaluation, the prompt also includes some in-context examples of answering multiple choice questions correctly. On the forget set, we refer to the accuracy of the model answering with the ground truth as the *MCQ Forget Accuracy* (stronger methods should decrease *MCQ Forget Accuracy*), and the accuracy of the model answering with the new edited answer’s choice as the *MCQ Edit Accuracy* (stronger methods should increase *MCQ Edit Accuracy*).

Finally, we evaluate the models’ post-editing accuracy on MMLU (Hendrycks et al., 2021) as a proxy for general language understanding to measure unintended side effects.

3.1.1. SPORTS TASKS

Sports Prompting 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 for the editing loss in the first place. For the MCQ evaluation, we use choices of all four sports (football, baseball, basketball, and golf). We average accuracies over all models, for Sports-Athlete-Editing we average over editing both 16 and 64 facts, and for Full-Sports-Editing we average over editing Basketball, Baseball, and Football.

Sports Results As shown in Figure 2, our analysis reveals that editing employing FLU localization exhibits superior performance in forgetting the original information and adopting the edited information, across different prompt formats. As explained in Section 3.1, better editing should result in higher *MCQ Forget Error*, and higher *MCQ Edit Accuracy*.

Figure 3 shows that FLU localized models are the best on both fronts. The difference is especially significant in Figure 3a, where only FLU model edits generalize meaningfully to MCQ, exceeding other localization methods by more than

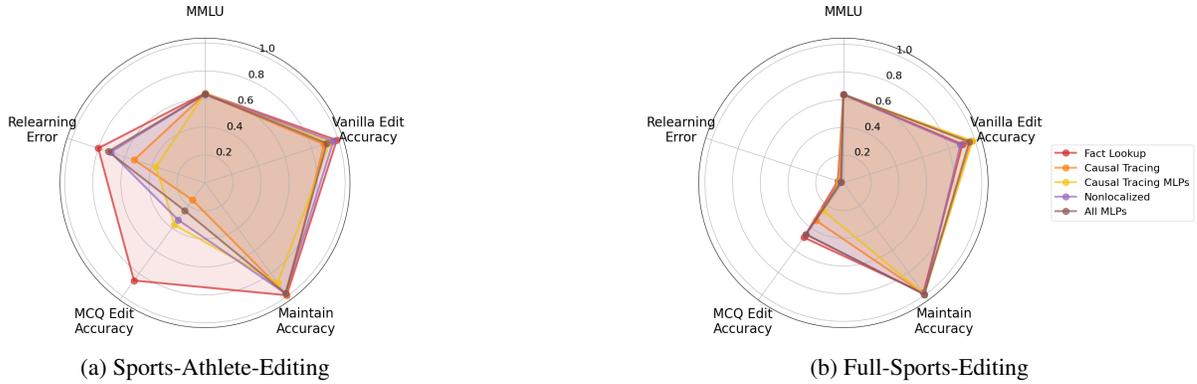


Figure 2. Spider plots illustrating the advantages of FLU for editing Sports across adversarial prompting and relearning evaluations, averaged over all three model types. **(Left)** The Sports-Athlete-Editing plot shows that FLU localization leads to editing that is the most robust against MCQ prompting and relearning. **(Right)** The plot shows that most localizations perform approximately equivalently in the Full-Sports-Editing task, with FLU localization slightly better for MCQ.

40% in MCQ Edit Accuracy.

A comprehensive comparison of all localization methods with multiple-choice prompting is available in Appendix A.7.2, further supporting our findings.

3.1.2. COUNTERFACT TASKS

CounterFact Prompting For CounterFact, we create an MCQ evaluation with four choices for every question, randomly ordering the true answer, the injected false answer, and two other question-specific LLM-generated incorrect answers. We also consider the original robustness and side effect evaluations from the Meng et al. (2023) dataset: the Paraphrase and Neighborhood facts accuracies. Answers of edited facts are meant to generalize to the Paraphrase evaluation, which phrases the fact in a different but equivalent way, so we report the *Paraphrase Edit Accuracy* (stronger methods should increase Paraphrase edit accuracy). Editing should not generalize to the Neighborhood evaluation, which presents similar but unrelated facts, so we report the *Neighborhood Edit Error* which is lower if models incorrectly report the edited answer in unrelated facts.

We again also use an MCQ evaluation, where the choices consist of the true answer, the injected false answer, and two other question-specific LLM-generated incorrect answers. We note that Paraphrase and Neighborhood evaluations ask for the answer in the same original format, so they are more in-distribution than MCQ. We average accuracies over all models, and for CounterFact-Editing we average over editing both 16 and 64 facts.

CounterFact Results Similar to our sports facts evaluation, Figure 4 shows an overview of key performance metrics across localizations for CounterFact. We see that FLU localization outperforms the baselines, being the only lo-

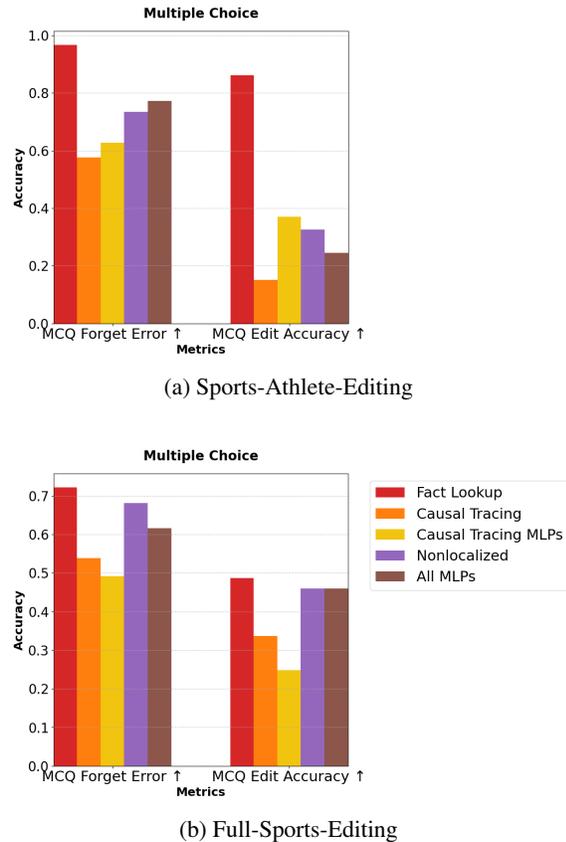


Figure 3. Bar charts showing results of MCQ evaluations, reporting both the forget error and edit accuracy when prompted with MCQ, averaged over three model types. For both **(a)** Sports-Athlete-Editing and **(b)** Full-Sports-Editing, FLU localization answers with the original answer the least (MCQ Forget Error) and answers with the edited answer most accurately (MCQ Edit Accuracy).

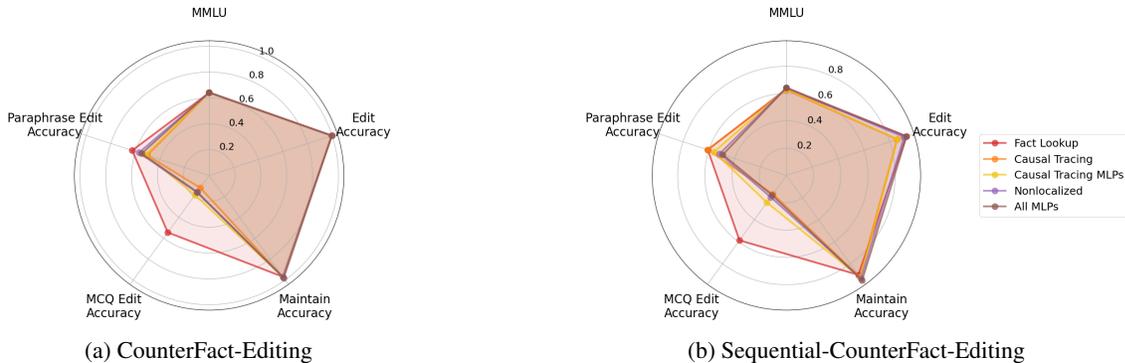


Figure 4. Spider plots illustrating the advantages of FLU for editing CounterFact across prompting evaluations, averaged over all three model types. (Left) The CounterFact-Editing plot shows that FLU localization leads to editing that is the most robust against MCQ prompting and Paraphrasing. (Right) The Sequential-CounterFact-Editing plot shows that FLU localization is the most robust against MCQ prompting.

calization to maintain MCQ edit performance. Figure 5 illustrates the robustness of FLU editing in the MCQ and Paraphrase evaluations. Edited models using FLU localization answer less frequently with the original, incorrect information (*MCQ Forget Error*) and more frequently provide the edited answer.

Furthermore, the *Neighborhood Edit Error* highlights that other localization methods exhibit slightly more pronounced side effects, inadvertently editing unintended, semantically similar facts.

Interestingly, sequential editing displays marginally greater robustness than nonsequential editing in MCQ when comparing between CounterFact-Editing and Sequential-CounterFact-Editing bars. This observation supports an approach that editing large sets of facts can be made more effective by partitioning the set and applying edits sequentially. We present results comparing all localizations from Section 2.2 across each prompt robustness evaluation in Appendix A.7.2, with consistent conclusions.

3.2. Adversarial Relearning Evaluation

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 a measure for the quality of editing. We replicate the methodology of Deeb & Roger (2024), splitting our forget sets in two independent halves, retraining with half of the ground truth labels, and evaluating on the other half. This methodology aims to discern whether the editing technique successfully removed the underlying factual association or merely obfuscated its direct retrieval while leaving it potentially susceptible to recovery when doing partial retraining.

We retrain with a rank-512 LoRA across all linear modules, with details available in Appendix A.7.3. We focus on the

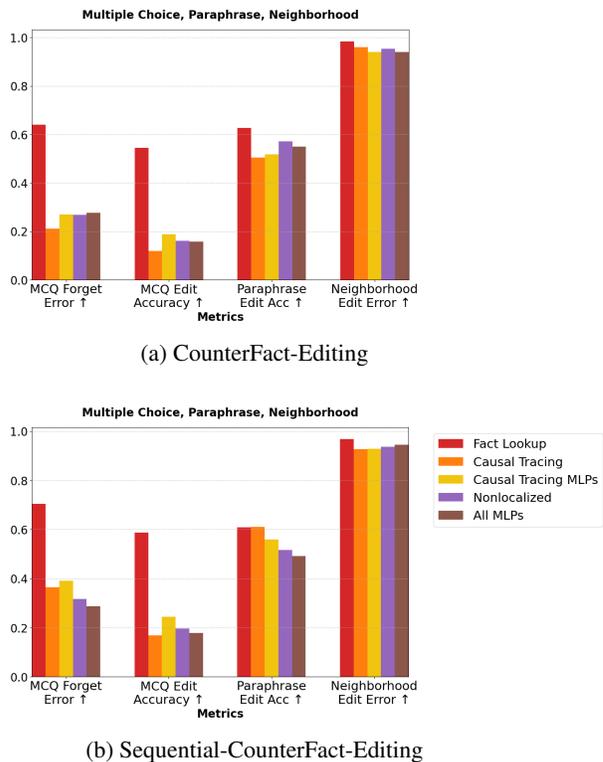


Figure 5. Bar charts showing results of MCQ, Paraphrase, and Neighborhood prompt evaluations, averaged over all three model types. For both (a) CounterFact-Editing and (b) Sequential-CounterFact-Editing, FLU localization has the most robust edit accuracy measured by MCQ and Paraphrase. FLU localization editing also does not incorrectly generalize to Neighborhood prompts. Sequential editing is slightly more robust than nonsequential editing in MCQ when comparing between CounterFact-Editing and Sequential-CounterFact-Editing bars.

Sports-Athlete-Editing task, as in the other tasks it was either too easy to relearn (Full-Sports-Editing) or too hard to relearn any performance (all CounterFact tasks) across all localizations. Relearning isn't a valid evaluation for Full-Sports-Editing because the facts are not independent, and models should reasonably generalize from relearning on half the basketball athletes to correctly answering the other half of the basketball athletes.

Sports Results Our adversarial relearning experiments, as depicted in Figure 6, reveal that retraining on a subset of the original “forgotten” data can recover a significant portion, as much as 63%, of the supposedly forgotten information when using OT methods like Causal Tracing and Causal Tracing MLPs. This suggests that these methods may simply mask direct retrieval of this information, leaving the model susceptible to this information recovery through retraining. In contrast, FLU localization exhibits greater resilience to such adversarial relearning, with only about 20% of the forgotten information recovered. This indicates that FLU localization may be more effective in targeting and removing the underlying knowledge, making it harder to recover through retraining.

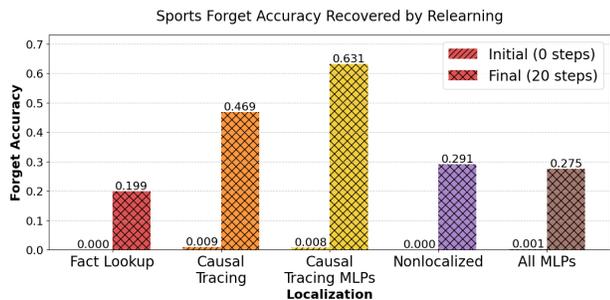


Figure 6. Relearning recovers the least accuracy on the forget set using FLU localizations. Relearning recovers significant accuracy on the original forget set in OT localizations (Causal Tracing and Causal Tracing MLPs).

We show results from Counterfact and from all localizations in Appendix A.7.3.

3.3. Latent Knowledge Analysis

In this section, we provide more evidence of our hypothesis that FLU unlearning targets the source of intermediate latent knowledge. We analyze the Sports-Athlete-Editing task again here because the ground truth and the edited answers vary between one of only three possibilities.

We train logistic regression models (probes) (Alain & Bengio, 2018) on prompt activations following every model layer to predict the correct ground truth sport on the maintained set of athletes. This is possible since there are only three possible sports, so we can train binary classification

probes for each sport and take the maximum classification over the sports. This discovers internal representations of the sport the model believes the answer to be: then, we apply these probes on the activations of the forget set of athletes.

We present graphs averaged over models and over 16 and 64 facts in Appendix A.7.4. We also present probing graphs for each individual model and all localizations in Appendix A.7.4. We demonstrate that probes on FLU consistently predict the forget sport less and the edit sport more than any other localization.

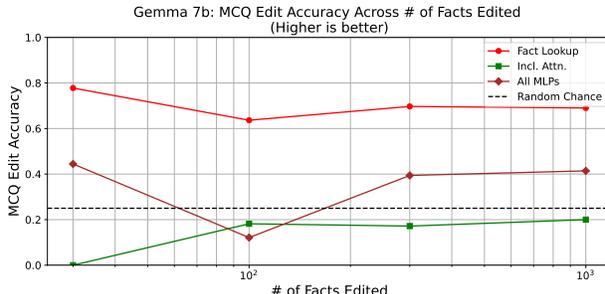


Figure 7. MCQ edit accuracy for Gemma-7B as number of facts edited is increased to 1000 facts. FLU localization is the only localization that preserves MCQ edit accuracy.

3.4. Fact Editing at Scale

We now examine the efficacy of our localization technique when editing between 30 to 1000 facts at once. This section also tests the use of an automated method of our “manual” FLU analysis. We automate our CounterFact analysis from Appendix A.2.1 by simply restricting to MLPs that have more than 2 standard deviation impact on the logit difference. We also now calculate a localization for each fact instead of taking the average across a batch of facts. We improve efficiency and avoid having to edit individual facts sequentially by batching facts with coincidentally the same localization during the fact editing process. To save computational resources, we perform this analysis on the Gemma-7B model and only compare our localization with the best baseline (all MLPs). In addition, we also test the effect of allowing attention heads to be part of the localization.

We see in Figure 7 that FLU localization preserves MCQ edit accuracy while the baselines fall to random chance levels. Figure 9 shows that FLU localization also has the highest forget error across edited fact amounts. Finally, we also provide latent knowledge analysis results at scale in Figure 8, showing that the FLU localization prevents latent knowledge extraction from being useful, and that including attention heads in the localization yields poor results towards the last layers of the model. We believe these results provide evidence that our methodology works well at scale.

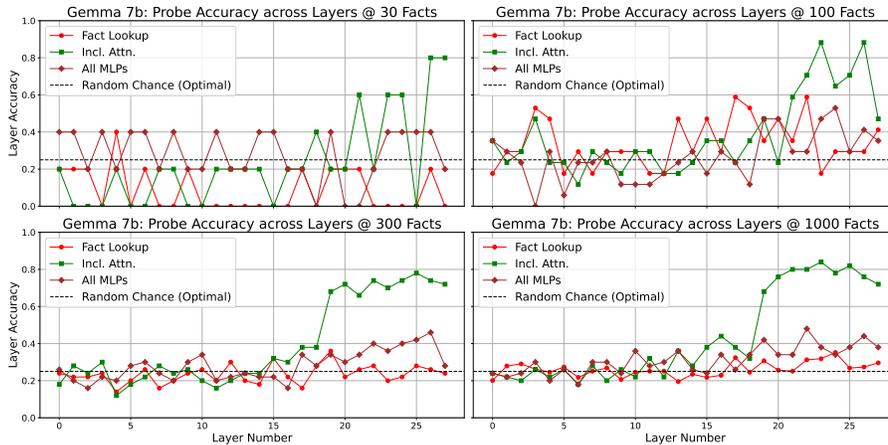


Figure 8. Latent knowledge analysis across layers for Gemma-7B when editing varying amounts of CounterFact facts. FLU localization and all MLPs maintain a random probe accuracy, resisting latent knowledge extraction unlike when including attention heads as part of the localization.

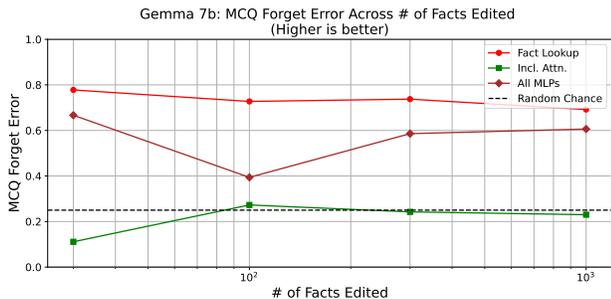


Figure 9. MCQ forget error for Gemma-7B when editing varying amounts of CounterFact facts. FLU localization maintains a better forget error than the baselines.

3.5. The Role of Parameter Count

In this section, we perform weight-masking involving training 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), to quantify the size of edits with different localizations and investigate which factual mechanisms are targeted when editing with different localizations. In this case, no weight updates are being performed. Rather, the mask turns a subset of the weights to zero.

Controlling for Parameter Count Although we already standardize the number of trainable parameters in most localizations, we additionally investigate if FLU editing is better than other localization techniques when controlling for the exact number of parameters that are masked. We perform weight masking on the Sports-Unlearning, Sports-Athlete-Editing, and CounterFact-Editing tasks. Detailed results are reported in Appendix A.4. We find that when con-

trolling for the size of the localization FLU is consistently more robust when subject to our suite of evaluations.

Other Localizations Affect the Extraction Mechanism

After training weight masks, we analyze the proportion of each mechanism (fact lookup, attribution extraction) that is masked by each localization’s weight mask in Appendix A.5. 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 fact lookup mechanisms.

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 editing robustness through localized fine-tuning on the FLU mechanism.

In Sections 3.3 and 3.5 and Appendix A.5, we provide evidence that OT and baseline approaches fail to be robust because they target extraction components, which fails to generalize and does not target the source of knowledge in the model. In contrast, FLU mechanistic understanding allows us to target editing at the sites where knowledge is sourced, which robustly prevents that information from entering the latent stream in any format.

Impact Statement

This paper advances the fields of interpretability and model editing, both of which are relevant for ensuring the safety, privacy, and fairness of models. We hope our methods help model developers responsibly manage harmful knowledge/behaviors.

References

- Alain, G. and Bengio, Y. Understanding intermediate layers using linear classifier probes, 2018.
- 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., Clark, J., McCandlish, S., Olah, C., Mann, B., and Kaplan, J. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.
- Bayazit, D., Foroutan, N., Chen, Z., Weiss, G., and Bosselut, A. Discovering knowledge-critical subnetworks in pretrained language models, 2023.
- Biderman, S., Schoelkopf, H., Anthony, Q., Bradley, H., O’Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., Skowron, A., Sutawika, L., and van der Wal, O. Pythia: A suite for analyzing large language models across training and scaling, 2023. URL <https://arxiv.org/abs/2304.01373>.
- Cao, Y. and Yang, J. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pp. 463–480. IEEE, 2015.
- Chughtai, B., Cooney, A., and Nanda, N. Summing up the facts: Additive mechanisms behind factual recall in llms, 2024.
- Conmy, A., Mavor-Parker, A. N., Lynch, A., Heimersheim, S., and Garriga-Alonso, A. Towards automated circuit discovery for mechanistic interpretability, 2023.
- Deeb, A. and Roger, F. Do unlearning methods remove information from language model weights?, 2024. URL <https://arxiv.org/abs/2410.08827>.
- Deng, J., Pang, S., Chen, Y., Xia, L., Bai, Y., Weng, H., and Xu, W. Sophon: Non-fine-tunable learning to restrain task transferability for pre-trained models. *arXiv preprint arXiv:2404.12699*, 2024.
- Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.
- Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. <https://transformer-circuits.pub/2021/framework/index.html>.
- Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., and Leahy, C. The pile: An 800gb dataset of diverse text for language modeling, 2020.
- Geva, M., Bastings, J., Filippova, K., and Globerson, A. Dissecting recall of factual associations in auto-regressive language models, 2023.
- Ginart, A., Guan, M. Y., Valiant, G., and Zou, J. Making ai forget you: Data deletion in machine learning, 2019.
- Goldowsky-Dill, N., MacLeod, C., Sato, L., and Arora, A. Localizing model behavior with path patching, 2023. URL <https://arxiv.org/abs/2304.05969>.
- Hase, P., Bansal, M., Kim, B., and Ghandeharioun, A. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models, 2023.
- Henderson, P., Mitchell, E., Manning, C., Jurafsky, D., and Finn, C. 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.
- Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding, 2021.
- Hong, Y., Yu, L., Yang, H., Ravfogel, S., and Geva, M. Intrinsic evaluation of unlearning using parametric knowledge traces, 2024. URL <https://arxiv.org/abs/2406.11614>.
- Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization, 2017.
- Lee, A., Bai, X., Pres, I., Wattenberg, M., Kummerfeld, J. K., and Mihalcea, R. A mechanistic understanding of alignment algorithms: A case study on dpo and toxicity. *arXiv preprint arXiv:2401.01967*, 2024.
- Lee, Y., Chen, A. S., Tajwar, F., Kumar, A., Yao, H., Liang, P., and Finn, C. Surgical fine-tuning improves adaptation to distribution shifts, 2023.

- Lermen, S., Rogers-Smith, C., and Ladish, J. Lora fine-tuning efficiently undoes safety training in llama 2-chat 70b, 2023.
- Lester, B., Al-Rfou, R., and Constant, N. The power of scale for parameter-efficient prompt tuning, 2021. URL <https://arxiv.org/abs/2104.08691>.
- Li, M., Davies, X., and Nadeau, M. Circuit breaking: Removing model behaviors with targeted ablation, 2024a.
- Li, N., Pan, A., Gopal, A., Yue, S., Berrios, D., Gatti, A., Li, J. D., Dombrowski, A.-K., Goel, S., Phan, L., Mukobi, G., Helm-Burger, N., Lababidi, R., Justen, L., Liu, A. B., Chen, M., Barrass, I., Zhang, O., Zhu, X., Tamirisa, R., Bharathi, B., Khoja, A., Zhao, Z., Herbert-Voss, A., Breuer, C. B., Marks, S., Patel, O., Zou, A., Mazeika, M., Wang, Z., Oswal, P., Lin, W., Hunt, A. A., Tienken-Harder, J., Shih, K. Y., Talley, K., Guan, J., Kaplan, R., Steneker, I., Campbell, D., Jokubaitis, B., Levinson, A., Wang, J., Qian, W., Karmakar, K. K., Basart, S., Fitz, S., Levine, M., Kumaraguru, P., Tupakula, U., Varadharajan, V., Wang, R., Shoshitaishvili, Y., Ba, J., Esvelt, K. M., Wang, A., and Hendrycks, D. The wmdp benchmark: Measuring and reducing malicious use with unlearning, 2024b.
- Lo, M., Cohen, S. B., and Barez, F. Large language models relearn removed concepts, 2024.
- Lynch, A., Guo, P., Ewart, A., Casper, S., and Hadfield-Menell, D. Eight methods to evaluate robust unlearning in llms, 2024.
- Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., Sakhaee, E., Li, N., Basart, S., Li, B., Forsyth, D., and Hendrycks, D. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal, 2024.
- Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in gpt, 2023.
- Nanda, N. Attribution patching: Activation patching at industrial scale, 2023. URL <https://www.neelnanda.io/mechanistic-interpretability/attribution-patching>.
- Nanda, N., Rajamanoharan, S., Kramár, J., and Shah, R. 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>.
- Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001. <https://distill.pub/2020/circuits/zoom-in>.
- Panigrahi, A., Saunshi, N., Zhao, H., and Arora, S. Task-specific skill localization in fine-tuned language models, 2023.
- Patil, V., Hase, P., and Bansal, M. Can sensitive information be deleted from llms? objectives for defending against extraction attacks, 2023.
- Polo, F. M., Weber, L., Choshen, L., Sun, Y., Xu, G., and Yurochkin, M. tinybenchmarks: evaluating llms with fewer examples, 2024.
- Syed, A., Rager, C., and Conmy, A. Attribution patching outperforms automated circuit discovery, 2023.
- Tamirisa, R., Bharathi, B., Phan, L., Zhou, A., Gatti, A., Suresh, T., Lin, M., Wang, J., Wang, R., Arel, R., Zou, A., Song, D., Li, B., Hendrycks, D., and Mazeika, M. Tamper-resistant safeguards for open-weight llms, 2024. URL <https://arxiv.org/abs/2408.00761>.
- Xhonneux, S., Dobre, D., Tang, J., Gidel, G., and Sridhar, D. In-context learning can re-learn forbidden tasks. *arXiv preprint arXiv:2402.05723*, 2024.
- Xu, H., Zhu, T., Zhang, L., Zhou, W., and Yu, P. S. Machine unlearning: A survey. *ACM Computing Surveys*, 56(1): 1–36, 2023.
- Yong, Z.-X., Menghini, C., and Bach, S. H. Low-resource languages jailbreak gpt-4, 2024.
- Yu, Q., Merullo, J., and Pavlick, E. Characterizing mechanisms for factual recall in language models, 2023.
- Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., Kolter, J. Z., and Hendrycks, D. Representation engineering: A top-down approach to ai transparency, 2023. URL <https://arxiv.org/abs/2310.01405>.
- Zou, A., Phan, L., Wang, J., Duenas, D., Lin, M., Andriushchenko, M., Wang, R., Kolter, Z., Fredrikson, M., and Hendrycks, D. Improving alignment and robustness with circuit breakers, 2024. URL <https://arxiv.org/abs/2406.04313>.

A. Appendix

A.1. Sports Unlearning Results

For unlearning on the Sports-Unlearning task, 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 L_{forget} we use is the $\log(1 - p)$ measure (where p is the probability of the correct sport) from Mazeika et al. (2024) due to its empirical stability and fewer side effects: vanilla gradient ascent more strongly incentivizes the model to have significantly lower logprobs than wouldn't be encountered in a model that has not been trained on the factual association, and it detracts from model maintenance of L_{retain} and L_{SFT} .

We present results on using various localizations on the Sports-Unlearning task in Table 1. The FLU localization allows unlearning to be more robust to the MCQ prompt format while maintaining performance on the MMLU dataset.

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

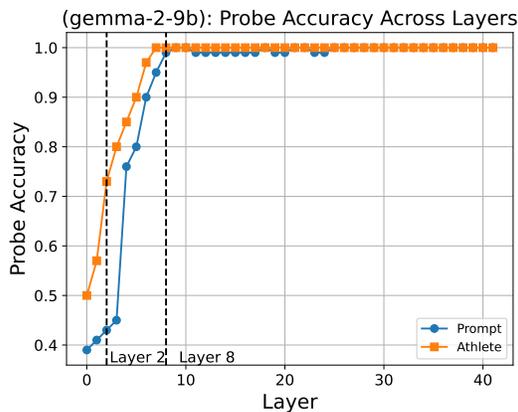
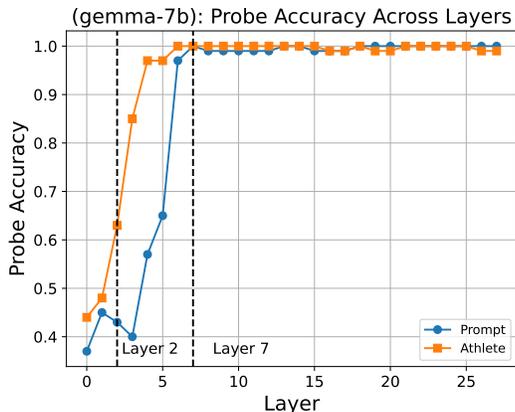
LOCALIZATION	FORGET ↓	RETAIN ↑	MCQ ↓	MMLU ↑
ATTRIB. PATCHING	0.000	1.000	0.767	0.602
CAUSAL TRACING	0.201	0.998	0.849	0.611
FLU	0.002	0.995	0.110	0.613
RANDOM	0.952	0.980	0.822	0.612
ALL-MLPs	0.000	0.994	0.279	0.606
NONLOCALIZED	0.000	0.985	0.196	0.595

A.2. FLU Interpretability Analysis

A.2.1. SPORTS FACTS

We redo analysis from Nanda et al. (2023) on Gemma-7B, Gemma-2-9B, and Llama-3-8B. 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 in Gemma-7B and 2 through 8 for Gemma-2-9B and Llama-3-8B (Figure 10).

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. 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.



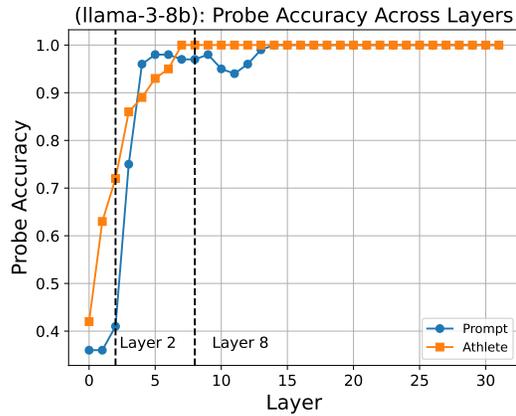


Figure 10. Accuracy of probes predicting the correct sport across layers for different models.

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 11. 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 12). 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 provide the logit differences for all 64 facts along with just the first 16 facts, and see that the logit differences are similar across the dataset splits. We take the MLPs with the highest change (> 0.02) and include them in our FLU localization of CounterFact.

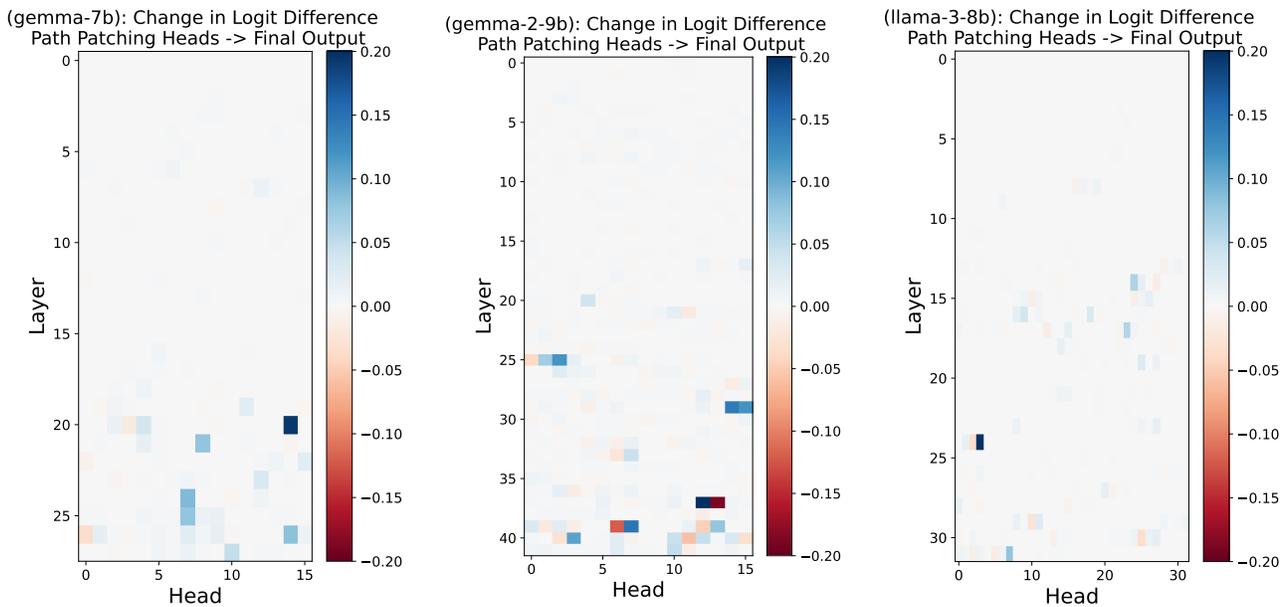


Figure 11. Change in logit difference when patching attention heads to the output.

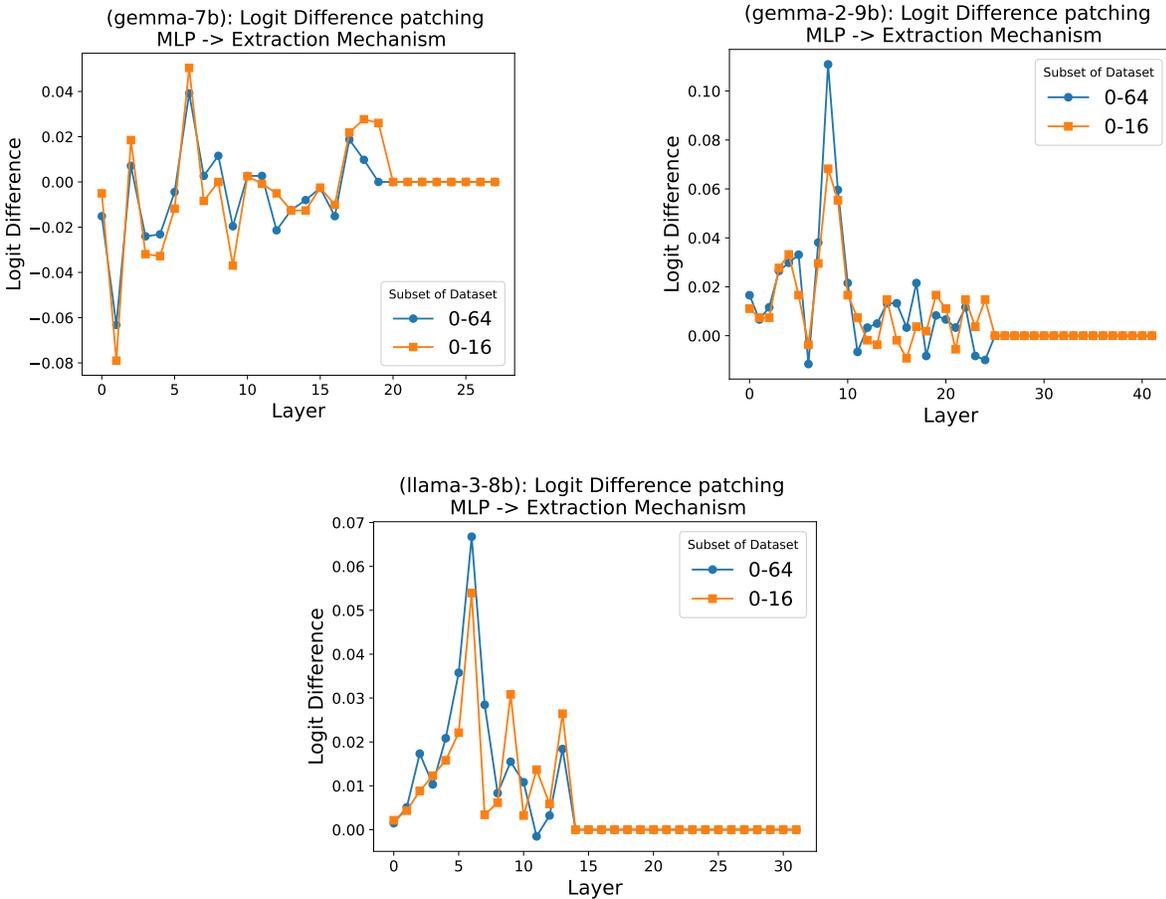


Figure 12. Logit difference when patching MLPs to the extraction mechanism found above for different models.

A.3. OT Selected Components

What MLPs do the automated OT methods localize? We explore the attribution scores of the automated localization methods (causal tracing and attribution patching) on the MLPs to see if automated localization methods can detect the FLU mechanism. In Figures 13 and 16, for Gemma-7B, we see that both CT and AP localizations target the later layer MLPs instead of the FLU mechanism (Figure 13).

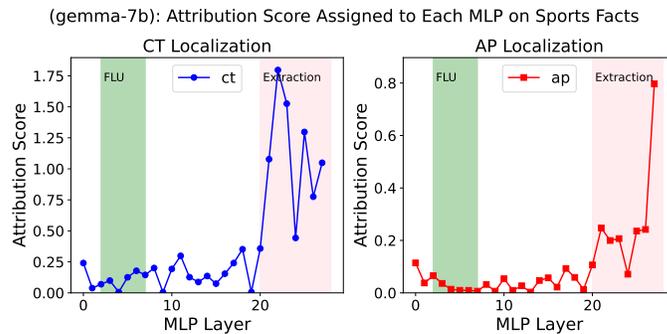


Figure 13. Attribution scores on MLPs on sports facts for Gemma-7B.

For AP localization, this trend continues with Gemma-2-9B (Figure 14) and Llama-3-8B (Figure 15). However, CT

localization does highlight some of the early layer MLPs that are in the FLU mechanism, especially for Gemma-2-9B.

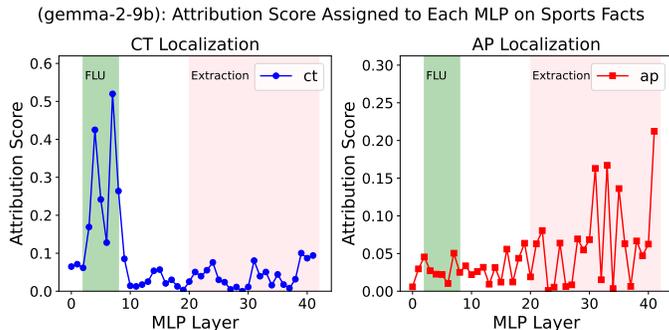


Figure 14. Attribution scores on MLPs on sports facts for Gemma-2-9B.

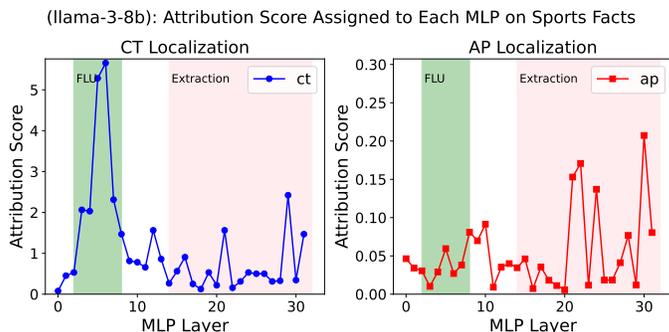


Figure 15. Attribution scores on MLPs on sports facts for Llama-3-8B.

We repeat this analysis on CounterFact in Figures 16 to 18. Again we see AP localizations in particular assign higher scores to later-layer MLPs, and CT only highlights FLU components on Gemma-2-9B, localizing other extraction layers on the other models.

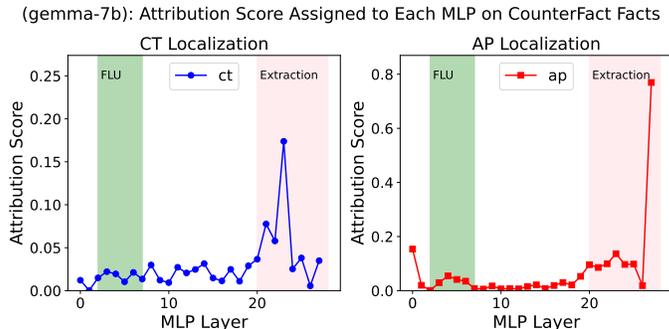


Figure 16. Attribution scores on MLPs on CounterFact facts for Gemma-7B.

A.4. 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. Our loss function $L = \lambda_1 L_{\text{forget}} + \lambda_2 L_{\text{retain}} + \lambda_3 L_{\text{SFT}} + \lambda_4 L_{\text{reg}}$ now includes an L1 regularization term to control the sparsity. 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. “Manual Interp” refers to the FLU localization technique for all the following results in this section.

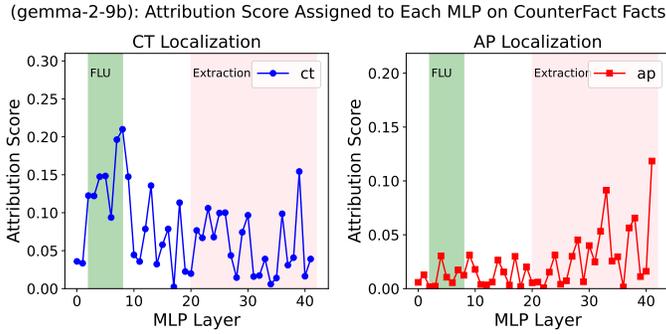


Figure 17. Attribution scores on MLPs on CounterFact facts for Gemma-2-9B.

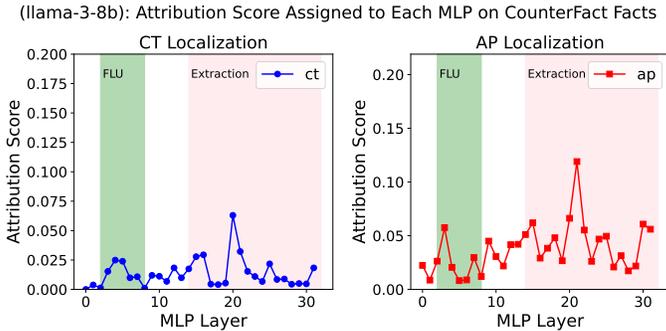


Figure 18. Attribution scores on MLPs on CounterFact facts for Llama-3-8B.

A.4.1. UNLEARNING SPORTS

We show standard evaluations across a sweep of discretization thresholds, which directly corresponds to the size of the model edit. Figure 19 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 19, right).

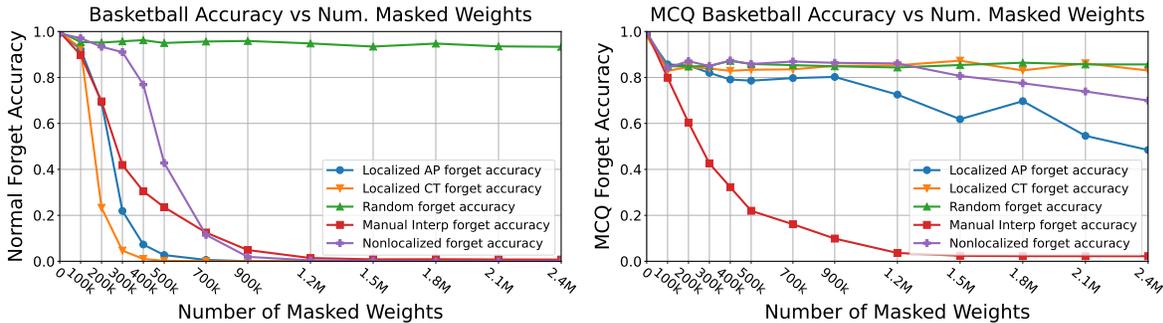


Figure 19. (Left) Testing the models' unlearning of basketball athletes against the number of weights masked. (Right) Testing the models' unlearning of basketball athletes against the number of weights masked, in the MCQ prompt format.

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 20, left), Figure 20 (right) shows manual localization generalizes for minimizing loss of MMLU capabilities while unlearning sports facts in the MCQ format compared to the other methods.

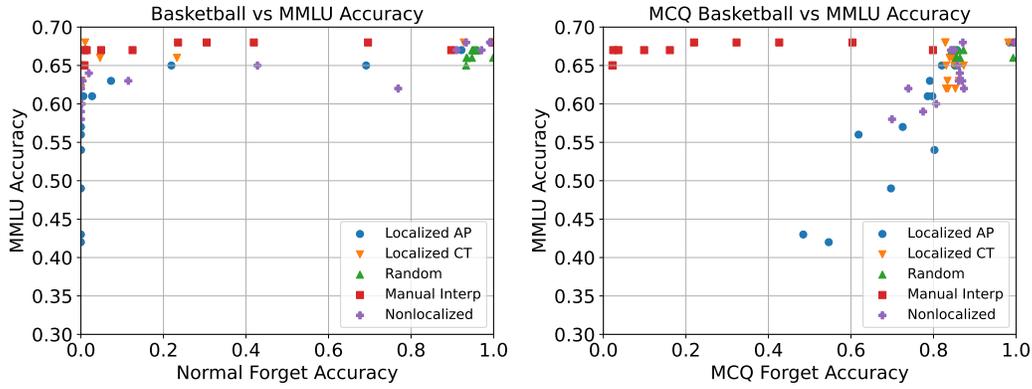


Figure 20. Unlearning basketball facts. **(Left)** Measuring MMLU and forget set performance across different discretization thresholds. **(Right)** Measuring MMLU and MCQ forget set performance across different discretization thresholds.

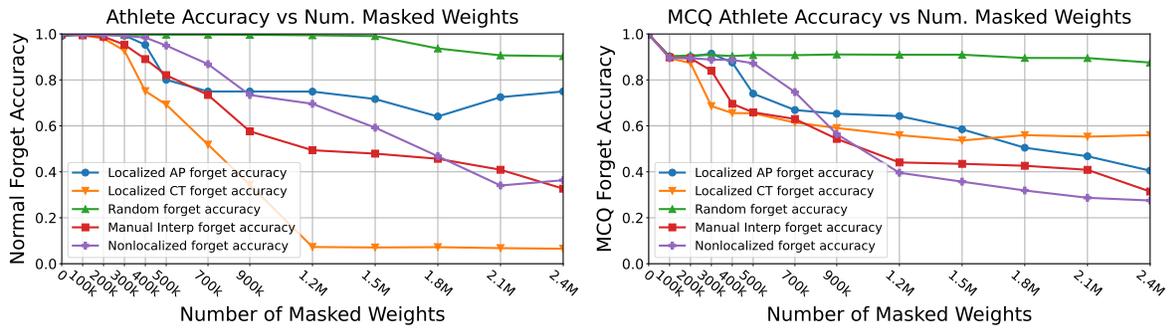


Figure 21. Editing subset of athletes. **(Left)** Measuring accuracy on the forget set. **(Right)** Measuring accuracy on the forget set in the MCQ prompt format.

A.4.2. EDITING ATHLETES

For editing the subset of athletes, Figure 21 shows that causal tracing localization causes the model to have 0% accuracy on the forget set, and FLU 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 22).

Furthermore, no other method completely generalizes this unlearning to the MCQ prompt format (Figure 21), and manual localization remains superior in minimizing loss of capabilities while unlearning the athlete subset (Figure 22, right).

A.4.3. EDITING COUNTERFACT

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 23.

Additionally, we report a comparison of all localizations across discretization thresholds for normal and MCQ forget sets in Figure 24 and Figure 25. We see that FLU outperforms all other methods of localization in preserving maintain accuracy while decreasing forget accuracy.

We perform additional adversarial analysis of accuracies across different discretization thresholds. We report the "paraphrase" and "neighborhood" adversarial results in Figure 26 and Figure 27, but find no significant results.

A.5. 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

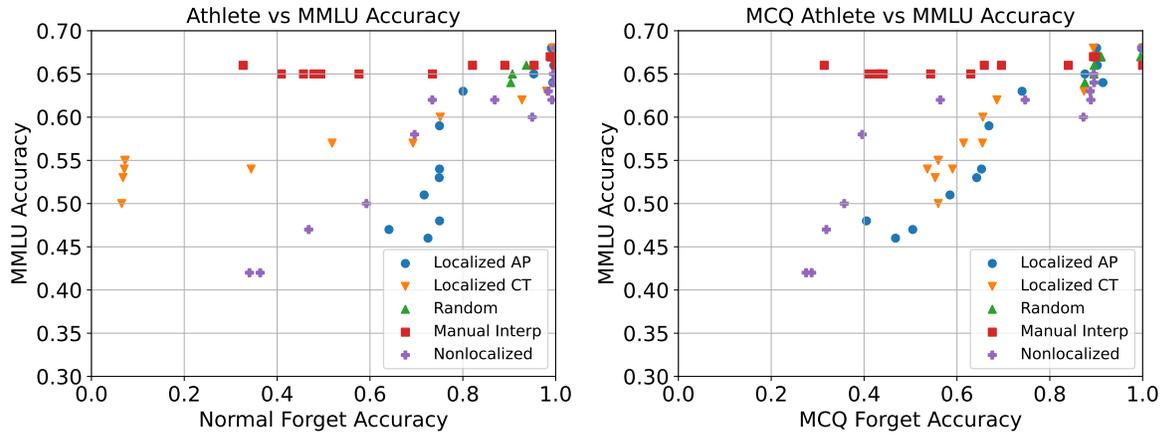


Figure 22. Editing subset of athletes. **(Left)** Measuring MMLU and forget set performance across different discretization thresholds. **(Right)** Measuring MMLU and MCQ forget set performance across different discretization thresholds.

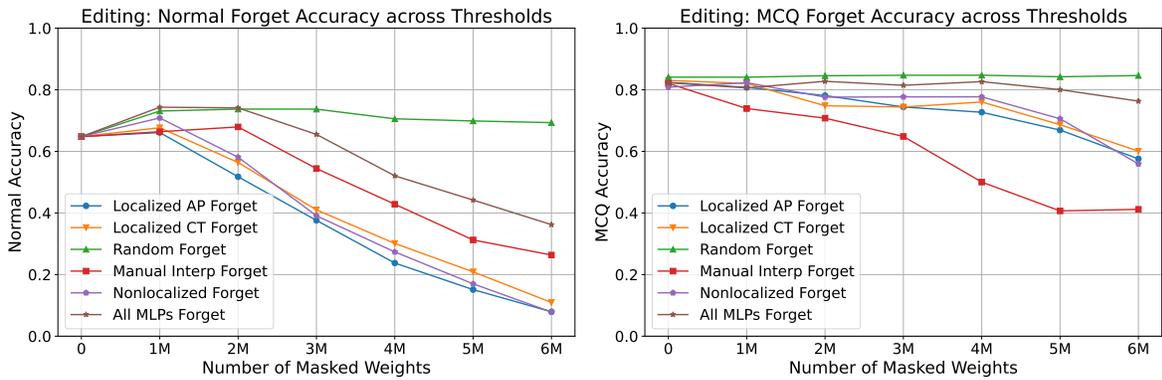


Figure 23. Editing CounterFact facts. **(Left)** Testing models accuracy on the normal forget set. **(Right)** Testing the models’ accuracy in the MCQ prompt format.

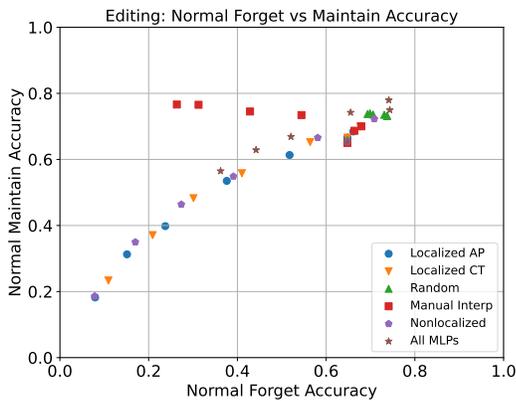


Figure 24. Accuracy on normal forget set vs on the maintain set across localizations and discretization thresholds.

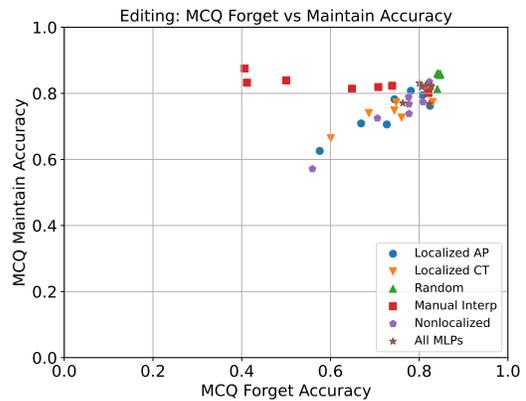


Figure 25. Accuracy on multiple choice input vs on the maintain set across localizations and discretization thresholds.

mechanisms.

First, in Table 2, we compare the parameter counts of the part of each mechanism that is present in each localization. Table 2 shows that causal tracing and attribution patching both have the potential to modify a considerable proportion of the

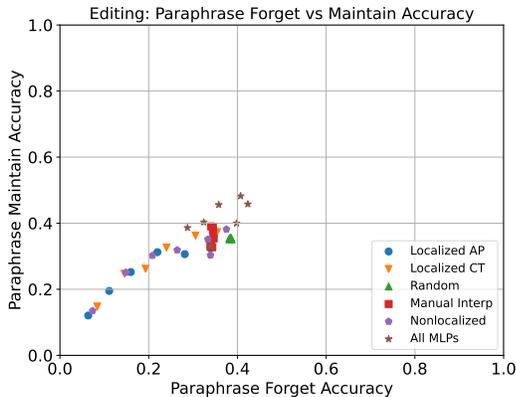


Figure 26. Accuracy on paraphrased input vs on the maintain set across localizations and discretization thresholds.

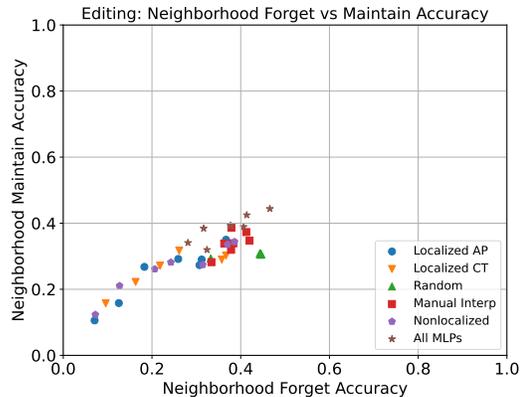


Figure 27. Accuracy on "neighborhood" input vs on the maintain set across localizations and discretization thresholds.

extraction heads and extraction MLPs.

Table 2. 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%)
FLU	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%)

Then, in Table 3, 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 3 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.

Table 3. 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%)
FLU	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%)

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 FLU localization is based on our discovered mechanisms, this does not serve as an evaluation of FLU (since by definition FLU localization will only localize the FLU mechanism), but rather only of causal tracing and attribution patching.

A.6. Hyperparameters

Across all tasks except Sequential-CounterFact-Editing and all models, we fine tune using 50 iterations of batch size 4 with 16 accumulation steps, using an AdamW optimizer (Kingma & Ba, 2017) with 0 weight decay and a cosine annealing scheduler. For Gemma-2-9b, we are forced to use an 8-bit optimizer to fit our training in the memory of 1 GPU. We find that the optimal learning rate is quite sensitive to the localization used and the edit task, so we first sweep over learning rates to find reasonable learning rates. We sweep over the learning rates of 2e-6, 5e-6, 1e-5, 2e-5, 5e-5, and 1e-4, training

models over 50 iterations with all λ s set to 1. We also tune the λ_1 parameter associated with the $L_{\text{injection}}$ cross entropy loss. We don't tune the other λ parameters because they are all maintenance losses, and setting them to 1 works sufficiently to maintain performance across almost all setups.

For the Sequential-CounterFact-Editing task, we use the same hyperparameters from the CounterFact-Editing task and we train for 100 total iterations rather than 50, using 25 for each subset of 16 facts. We choose 25 iterations because it balances between being half the number of iterations we typically use per subset of that size, and also double the number of steps overall as we use in CounterFact-Editing.

To avoid leaking evaluation information through this sweep process, we optimize learning rate for the objective of $(1 - \text{Forget Set Ground Truth Accuracy}) + \text{Forget Set Edit Accuracy} + \text{Maintain Set Ground Truth Accuracy} + \text{Pile Accuracy}$, avoiding any of our robustness metrics (one could view this sweep process as simply another part of the training process, since we only use train-time information). We run sweeps for Causal Tracing, Manual Fact Lookup, and No Localization. We then use the hyperparameters from Causal Tracing for Attribution Patching and Random (which all localize to MLPs and attention components), we use the hyperparameters from Fact Lookup for Random MLPs, Causal Tracing MLPs, and Attribution Patching MLPs (all localize to the same number of MLPs), and we use the hyperparameters from No Localization for All MLPs (which have the largest number of active parameters).

A.6.1. SAMPLE HYPERPARAMETER SWEEP

In Table 4, we show the full results of one sweep, optimizing learning rate for editing 16 facts from CounterFact. Especially for No Localization, some learning rates fail to edit in the correct answer with high accuracy, or fail to maintain accuracy on the maintain set. In Table 5, we see that editing results are not particularly sensitive to the coefficient used with the injection cross entropy loss.

Table 4. Gemma-7b learning rate sweep, editing 16 CounterFact facts.

	Pile Accuracy \uparrow	Forget Accuracy \downarrow	Edit Accuracy \uparrow	Maintained Accuracy \uparrow	Overall Score \uparrow
<i>FLU</i>					
LR 0.0001	0.488	0.000	1.000	0.698	3.186
LR 1e-05	0.513	0.000	1.000	0.975	3.488
LR 2e-05	0.542	0.000	1.000	0.950	3.492
LR 2e-06	0.499	0.007	0.961	0.869	3.323
LR 5e-05	0.520	0.000	1.000	0.822	3.342
LR 5e-06	0.528	0.000	0.999	0.980	3.507
<i>Localized CT</i>					
LR 0.0001	0.462	0.001	0.999	0.697	3.157
LR 1e-05	0.513	0.000	1.000	0.984	3.496
LR 2e-05	0.507	0.000	1.000	0.961	3.467
LR 2e-06	0.540	0.036	0.921	0.849	3.274
LR 5e-05	0.488	0.000	1.000	0.846	3.334
LR 5e-06	0.537	0.000	1.000	0.982	3.519
<i>Nonlocalized</i>					
LR 0.0001	0.062	0.032	0.557	0.094	1.680
LR 1e-05	0.520	0.000	1.000	0.892	3.412
LR 2e-05	0.479	0.001	0.998	0.807	3.284
LR 2e-06	0.529	0.000	1.000	0.982	3.510
LR 5e-05	0.046	0.034	0.710	0.092	1.815
LR 5e-06	0.536	0.000	1.000	0.988	3.524

A.6.2. ALL HYPERPARAMETERS USED

Table 6 has all learning rates used and Table 7 has all injection loss coefficients used.

A.7. Evaluation Details

A.7.1. DETAILS ON STANDARD PROMPT EVALUATIONS

We report standard metrics of Forget Error, Edit Accuracy, and Maintain Accuracy, in the same prompt format that the models were trained on. These metrics are optimized by the loss, so we expect all localizations to do almost perfectly on these evaluations. Figures 28 to 31 show that localizations perform approximately equivalently on these basic evaluations

Table 5. Gemma-7b inject loss coefficient sweep, editing 16 CounterFact facts.

	Pile Accuracy \uparrow	Forget Accuracy \downarrow	Edit Accuracy \uparrow	Maintained Accuracy \uparrow	Overall Score \uparrow
<i>FLU</i>					
FC 0.1	0.538	0.000	0.998	0.990	3.525
FC 0.2	0.510	0.000	0.999	0.981	3.491
FC 0.5	0.508	0.000	0.999	0.985	3.493
FC 1	0.510	0.000	0.999	0.973	3.483
FC 2	0.532	0.000	1.000	0.984	3.515
FC 5	0.534	0.000	1.000	0.985	3.518
<i>Localized CT</i>					
FC 0.1	0.532	0.004	0.995	0.983	3.506
FC 0.2	0.535	0.001	0.998	0.986	3.518
FC 0.5	0.524	0.000	0.999	0.974	3.497
FC 1	0.519	0.000	1.000	0.988	3.507
FC 2	0.536	0.000	1.000	0.979	3.514
FC 5	0.518	0.000	1.000	0.975	3.493
<i>Nonlocalized</i>					
FC 0.1	0.525	0.001	0.998	0.969	3.492
FC 0.2	0.562	0.000	0.999	0.982	3.543
FC 0.5	0.524	0.000	1.000	0.962	3.486
FC 1	0.531	0.000	1.000	0.980	3.511
FC 2	0.514	0.000	1.000	0.978	3.492
FC 5	0.528	0.000	1.000	0.975	3.503

Table 6. Optimal learning rates for different models, task types, and localizations.

Model	64 athletes to random sport	Basketball Athletes to Golf	16 CounterFact facts	64 CounterFact facts
Gemma				
FLU	1×10^{-5}	2×10^{-5}	5×10^{-6}	1×10^{-5}
Localized CT	1×10^{-5}	5×10^{-6}	5×10^{-6}	2×10^{-5}
Nonlocalized	2×10^{-6}	5×10^{-6}	5×10^{-6}	2×10^{-6}
Gemma 2				
FLU	1×10^{-5}	5×10^{-5}	5×10^{-5}	2×10^{-5}
Localized CT	5×10^{-5}	5×10^{-5}	5×10^{-5}	2×10^{-5}
Nonlocalized	2×10^{-5}	1×10^{-5}	5×10^{-6}	5×10^{-6}
Llama 3				
FLU	5×10^{-5}	5×10^{-5}	1×10^{-4}	5×10^{-5}
Localized CT	1×10^{-4}	5×10^{-5}	2×10^{-5}	5×10^{-5}
Nonlocalized	2×10^{-5}	1×10^{-5}	2×10^{-5}	1×10^{-5}

Table 7. Optimal inject loss coefficients for different models, task types, and localizations.

Model	64 athletes to random sport	Basketball Athletes to Golf	16 CounterFact facts	64 CounterFact facts
Gemma				
FLU	5.0	5.0	0.1	2.0
Localized CT	0.1	1.0	0.2	1.0
Nonlocalized	0.2	1.0	0.2	1.0
Gemma 2				
FLU	1.0	5.0	2.0	0.5
Localized CT	5.0	5.0	2.0	2.0
Nonlocalized	5.0	5.0	1.0	1.0
Llama 3				
FLU	5.0	0.2	1.0	0.2
Localized CT	2.0	1.0	0.1	2.0
Nonlocalized	2.0	0.1	2.0	0.2

across all tasks.

Mechanistic Unlearning

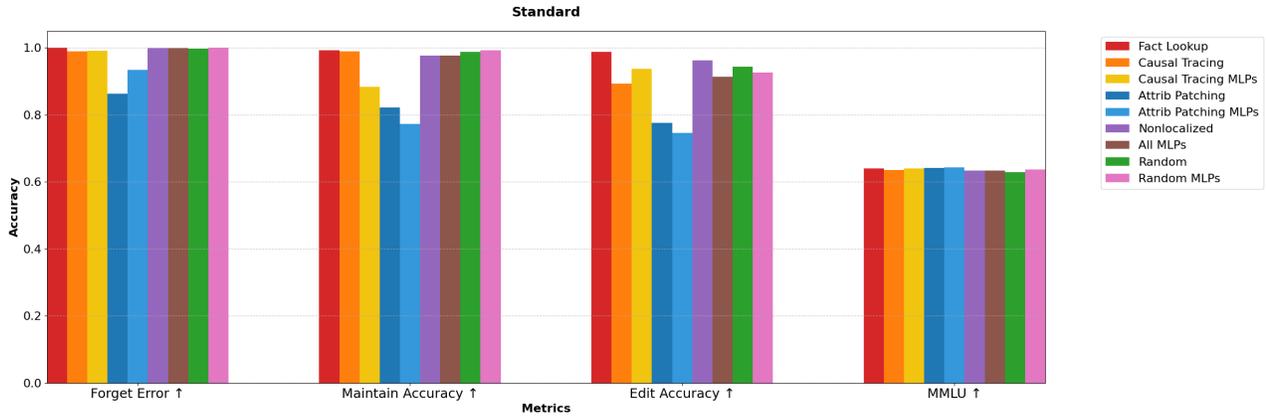


Figure 28. Standard Prompting results for Sports-Athlete-Editing, across all localizations.

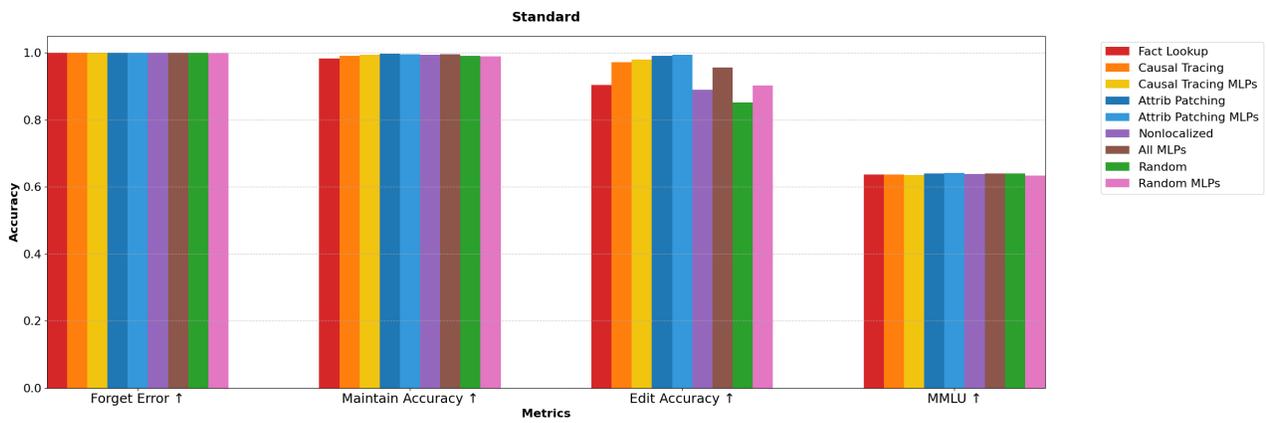


Figure 29. Standard Prompting results for Full-Sports-Editing, across all localizations.

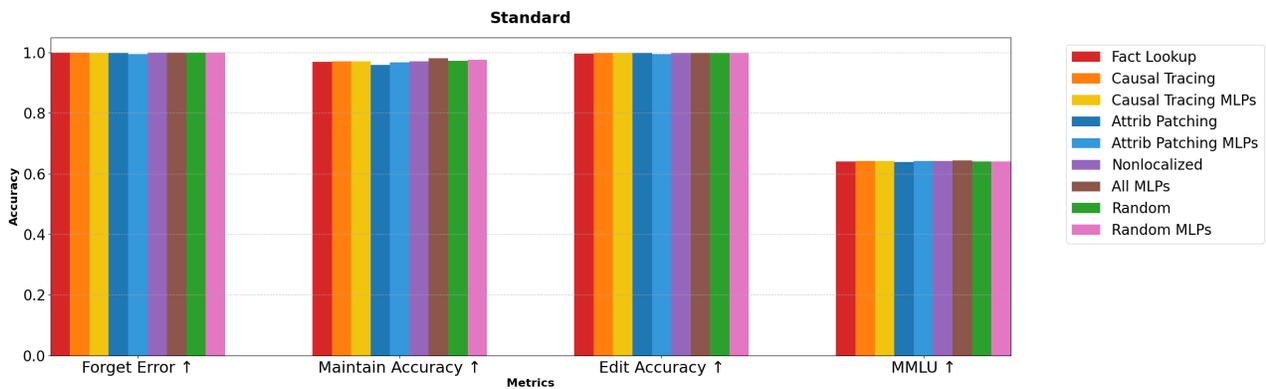


Figure 30. Standard Prompting results for CounterFact-Editing, across all localizations.

A.7.2. DETAILS ON ADVERSARIAL PROMPT EVALUATIONS

We report the adversarial prompt evaluations from Section 3.1 across all localizations. Figures 32 to 35 all show that FLU localization is more robust in MCQ compared to every other localization (significantly stronger for CounterFact). Figures 34 and 35 show that FLU localization is optimal in Paraphrase and Neighborhood in all cases except for Paraphrase compared to the Random localization in Sequential-CounterFact-Editing.

Mechanistic Unlearning

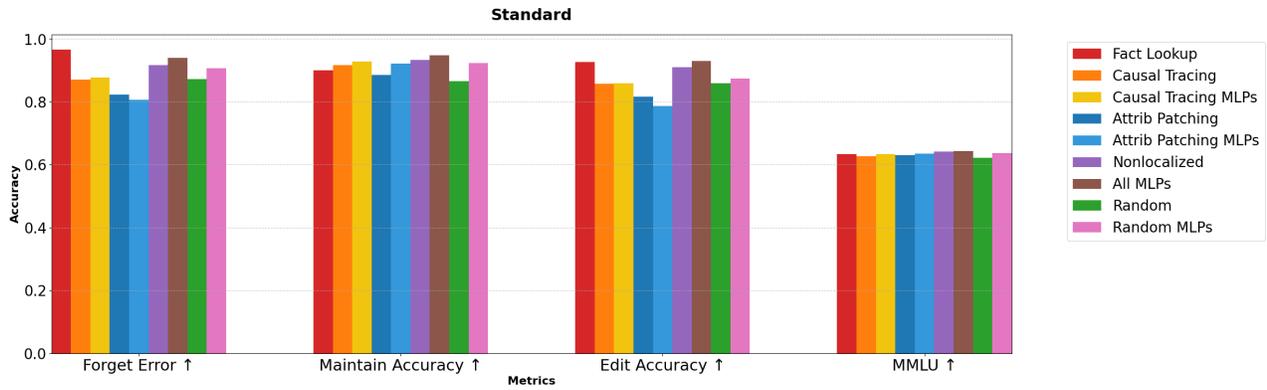


Figure 31. Standard Prompting results for Sequential-CounterFact-Editing, across all localizations.

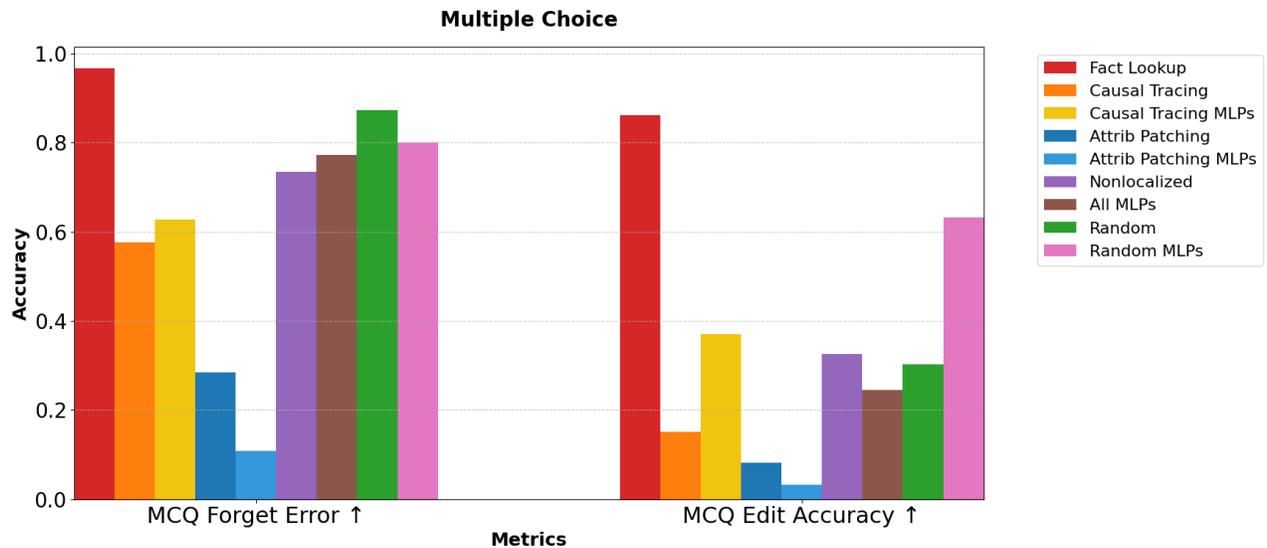


Figure 32. Adversarial Prompting results for Sports-Athlete-Editing, across all localizations.

A.7.3. DETAILS ON ADVERSARIAL RELEARNING

We retrain the model for 20 iterations with cross-entropy on half of the forget set (along with a standard retain and SFT loss), adding up all losses with loss coefficient 1.

We present relearning results for all localizations averaged over models. As shown in Figure 36, the FLU localization remains optimal, although the baselines of Nonlocalized, All MLPs, Random, and Random MLPs are competitive.

We also present relearning results on the other tasks. As mentioned in Section 3.2, since Full-Sports-Editing forget facts are not independent, we don't expect valid results from relearning. Thus, in Figure 37, we see that every localization regains 100% editing accuracy.

On CounterFact-Editing and Sequential-CounterFact-Editing, as shown in Figures 38 and 39, none of the localizations relearn more than 7% accuracy, suggesting adversarial relearning was not a sufficiently strong enough evaluation for these tasks. Regardless, FLU localization is either the most or second-most robust localization to relearning, although localizations don't differ by much.

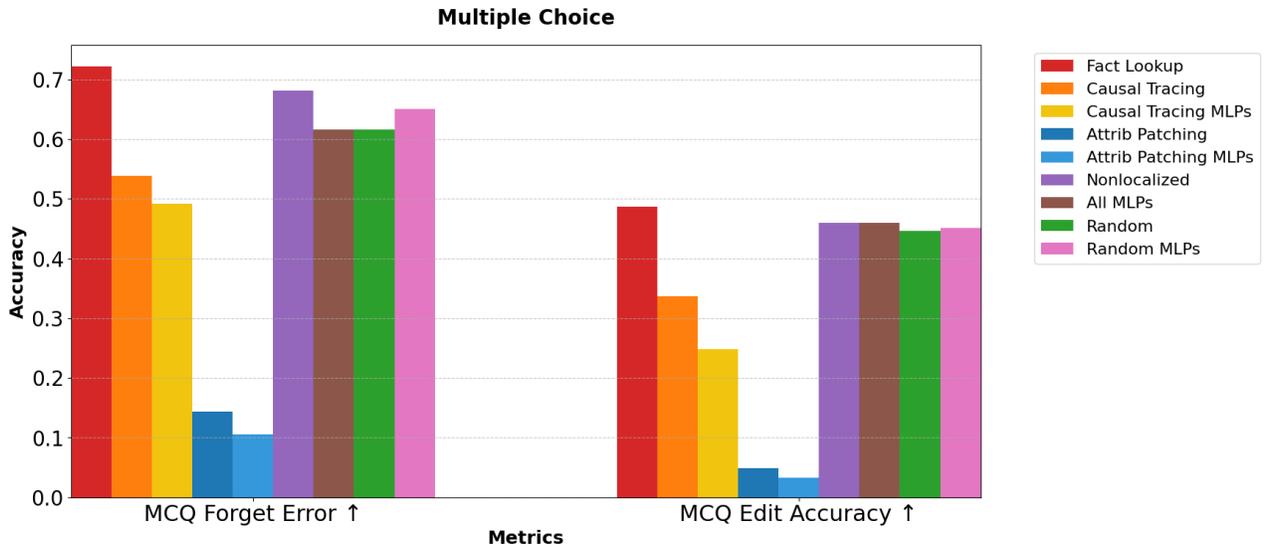


Figure 33. Adversarial Prompting results for Full-Sports-Editing, across all localizations.

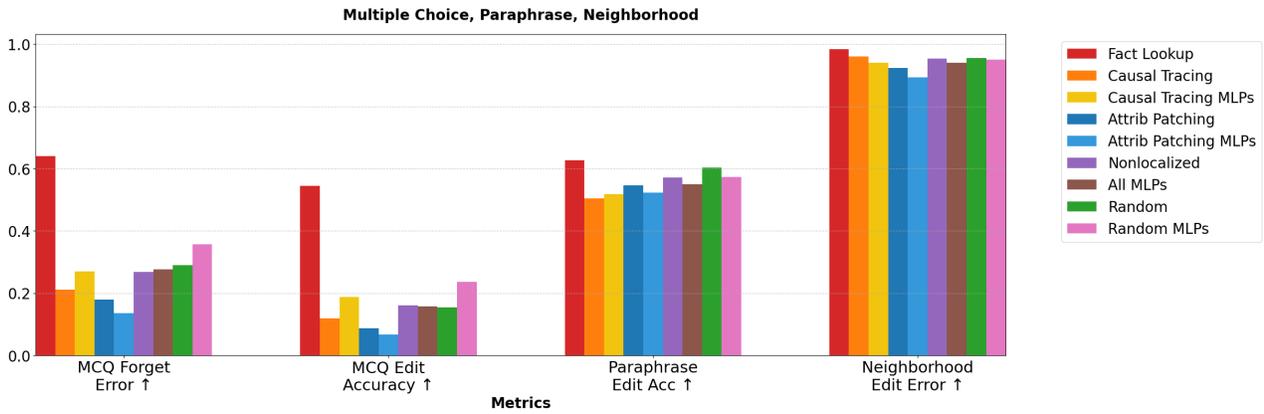


Figure 34. Adversarial Prompting results for CounterFact-Editing, across all localizations.

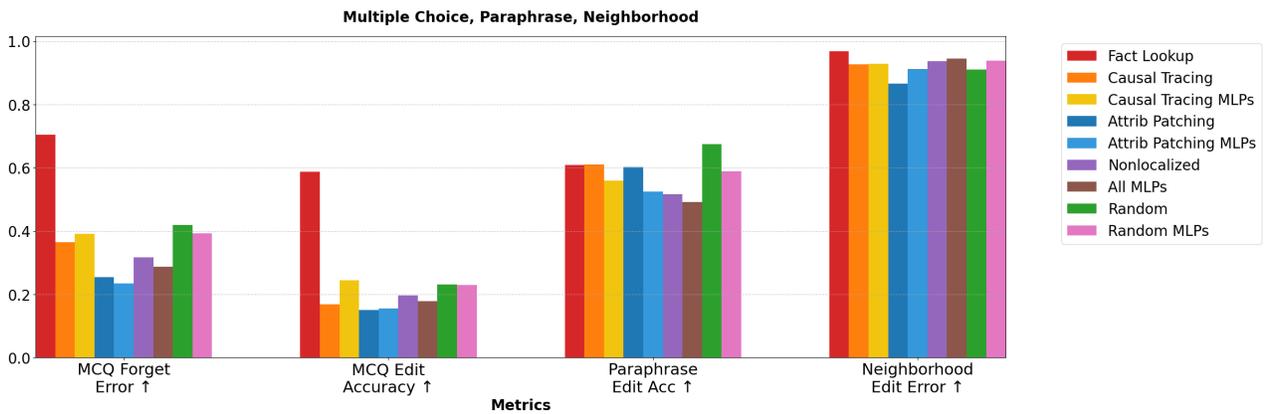


Figure 35. Adversarial Prompting results for Sequential-CounterFact-Editing, across all localizations.

Mechanistic Unlearning

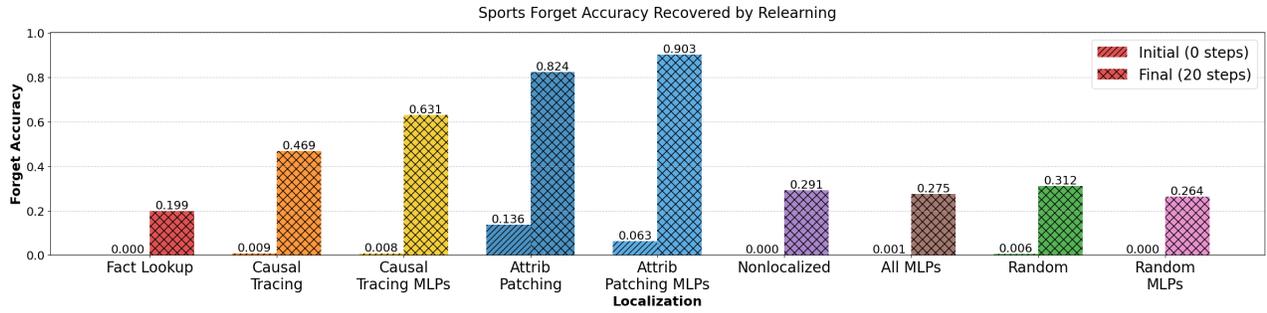


Figure 36. Relearning results for Sports-Athlete-Editing, across all localizations.

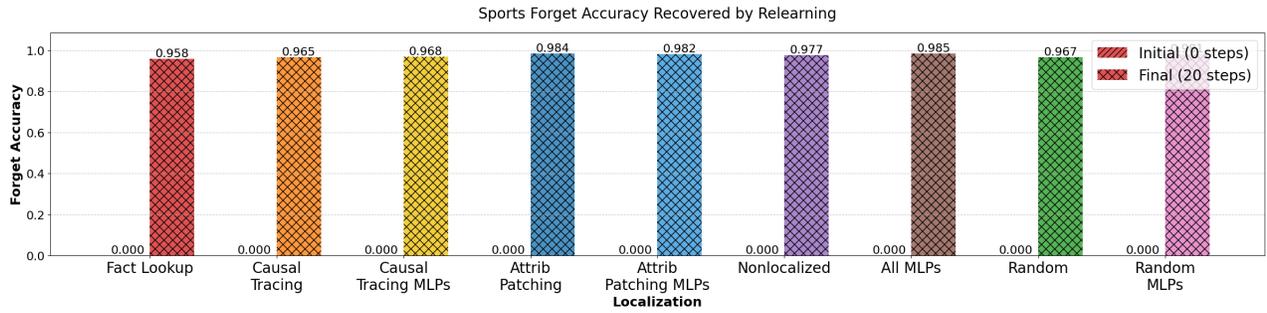


Figure 37. Relearning results for Full-Sports-Editing.

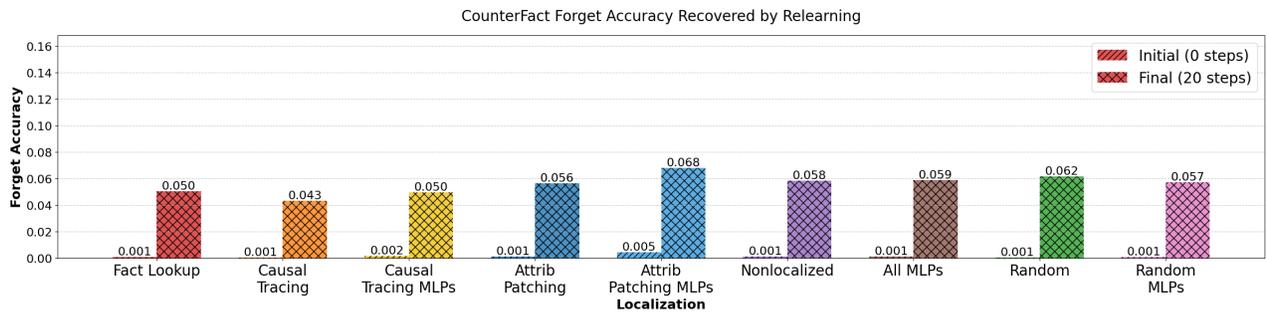


Figure 38. Relearning results for CounterFact-Editing.

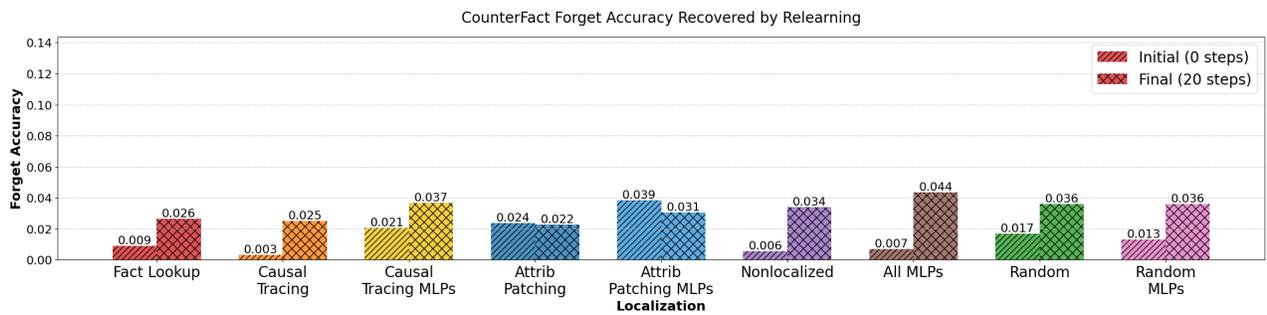


Figure 39. Relearning results for Sequential-CounterFact-Editing.

A.7.4. DETAILS ON LATENT KNOWLEDGE

In Figure 40, the probes on FLU consistently predict the forget sport less and the edit sport more than in any other localization, especially in early layers. Furthermore, the FLU probe classifications for the most part monotonically converge from their nonzero starting accuracy to 0 (for forget accuracy) and 1 (for edit accuracy).

Every other localization has much higher peak probe classification forget accuracy in the early layers, especially the OT localizations which have peak classification forget accuracy of almost 100%. This strongly suggests that these models still significantly represent the ground truth answer rather than the edit answer in early layers.

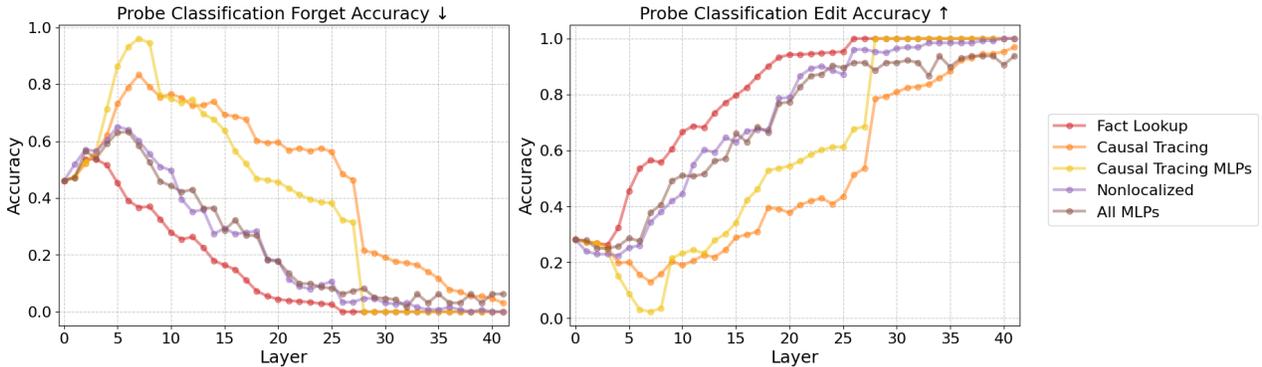


Figure 40. Linear probes applied to the forget set, classifying model activations after various layers. (Left) The line graph shows that some localizations still represent almost completely correct forget set knowledge in early layers, especially OT, while FLU localizations represent this original knowledge the least. (Right) The line graph shows that FLU localizations represent the edited rather than original answer earlier and more consistently throughout layers than any other localization.

We present the probing classification accuracies for the three models separately here, as well as for all localizations we previously left out.

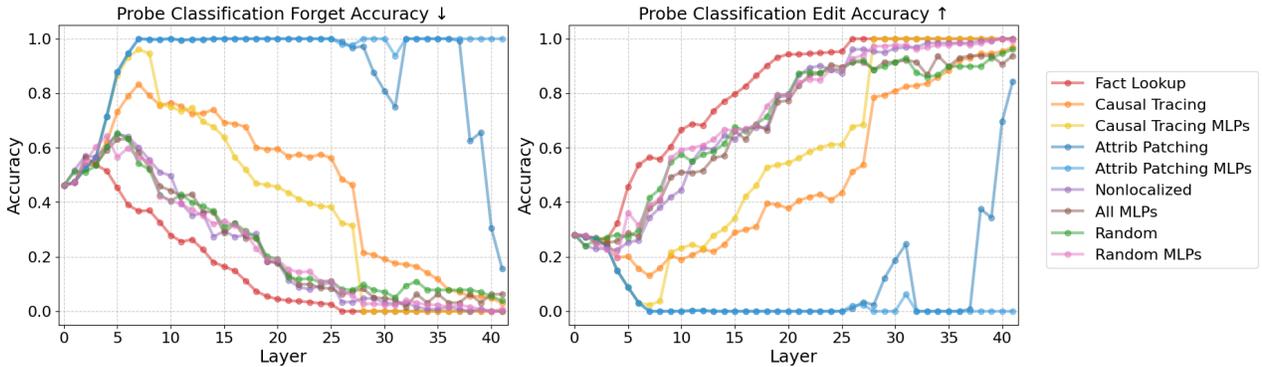


Figure 41. Linear probes applied to the forget set across all models, classifying model activations after various layers.

In Gemma-7b and Llama-3-8b, FLU probing is the most monotonic and the best in the early layers, either steadily decreasing to 0 for forget accuracy or increasing to 1 for edit accuracy, with the least extreme peaks. In Gemma-2-9b, the Nonlocalized, All MLPs, and Random MLPs baselines are competitive with FLU. The other OT localization, Attribution Patching, has 100% probing forget accuracy across many layers, suggesting it represents the ground truth answer very clearly.

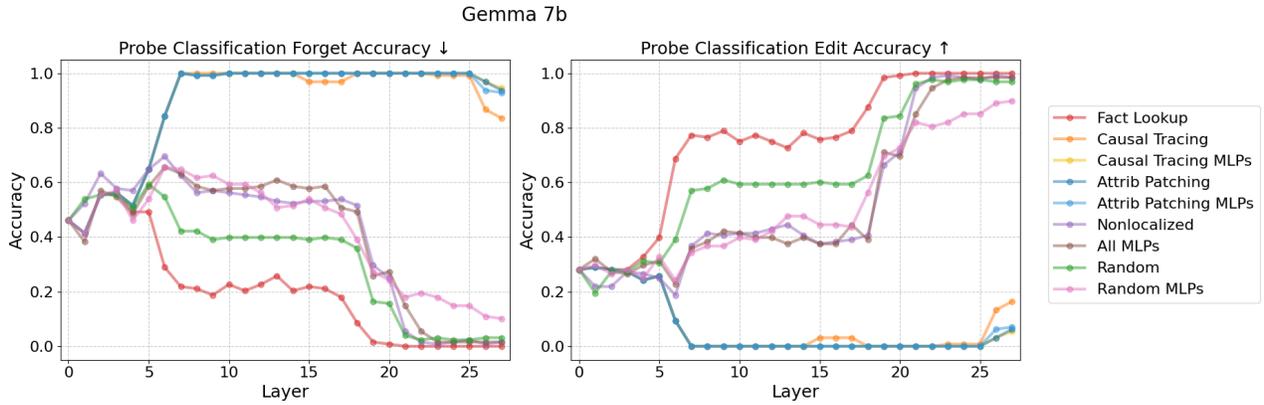


Figure 42. Linear probes applied to the forget set on Gemma-7B with 28 layers.

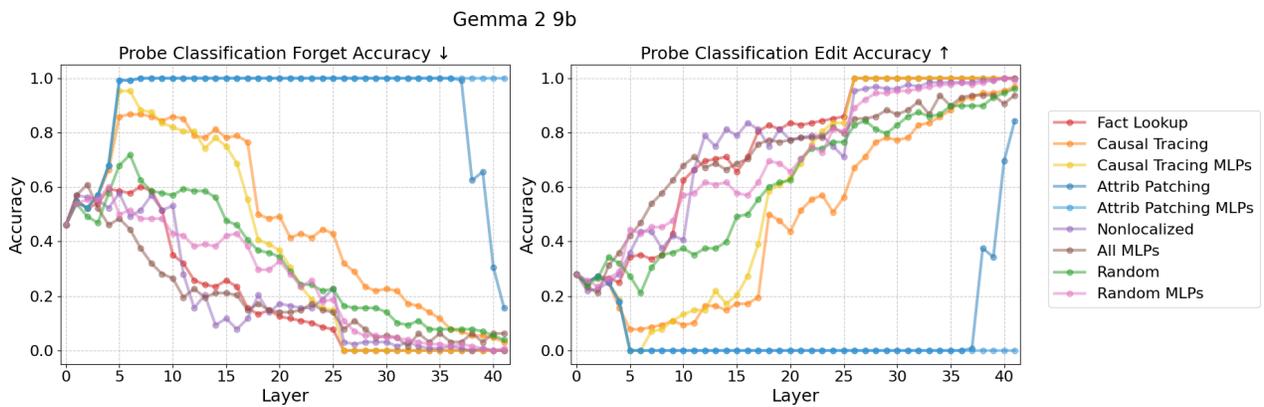


Figure 43. Linear probes applied to the forget set on Gemma-2-9b with 42 layers.

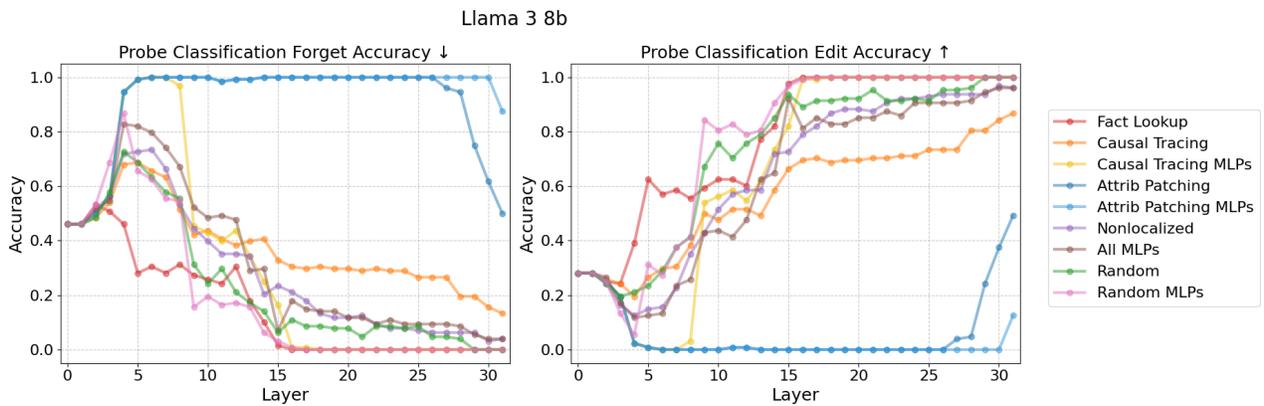


Figure 44. Linear probes applied to the forget set on Llama-3-8b with 32 layers.

A.8. Soft Prompt Evaluations

Because many localizations seem to be weak to prompting schemes, we attempt a simple adaptive attack of soft prompts, where we optimize the continuous embeddings at the end of the prompt to recover the correct answer on half of our forget set. We then evaluate the model’s performance on the other half, with this soft prompt in place (Lester et al., 2021). We average over evaluations from four soft prompts. Soft prompt evaluations can be considered to be a more narrow form of

few-shot finetuning, that is closer to searching for prompts that recover the model’s knowledge.

We find limited soft prompt success: across most tasks and models, we don’t recover much held-out forget set accuracy. On the Sports-Athlete-Editing, CounterFact-Editing, and Sequential-CounterFact-Editing tasks, Figures 45, 47 and 48 show that all localizations don’t significantly improve in Forget Accuracy over random chance, or are about equal between localizations, after soft prompts are applied. In Figure 46, specifically for Gemma-2 on Sports-Athlete-Editing we see some reasonable results with softprompts that are able to recover over 60% Forget Accuracy on OT localizations, while FLU, Nonlocalized, and All MLPs remain under 40% Forget Accuracy.

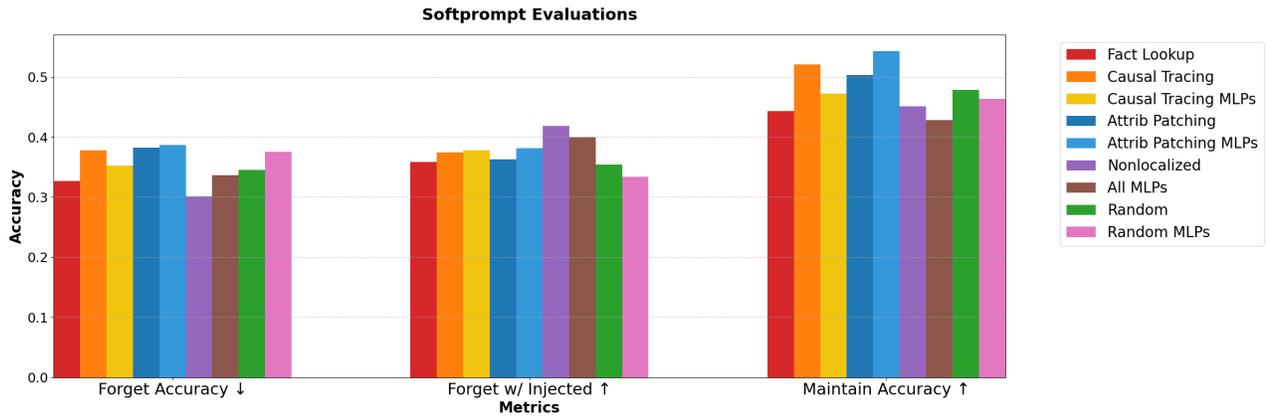


Figure 45. Metrics with soft prompts applied for Sports-Athlete-Editing, averaged over all models.

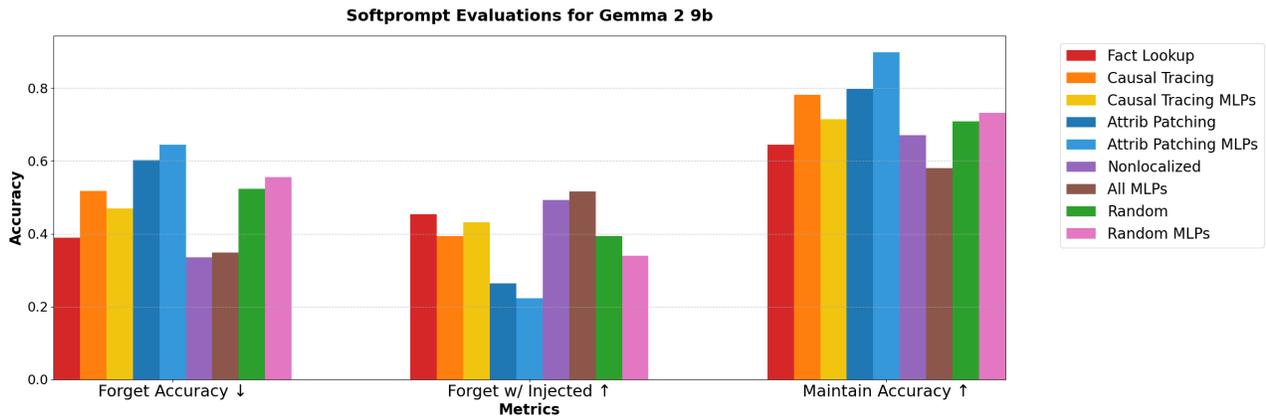


Figure 46. Metrics with soft prompts applied for Sports-Athlete-Editing for Gemma-2-9b.

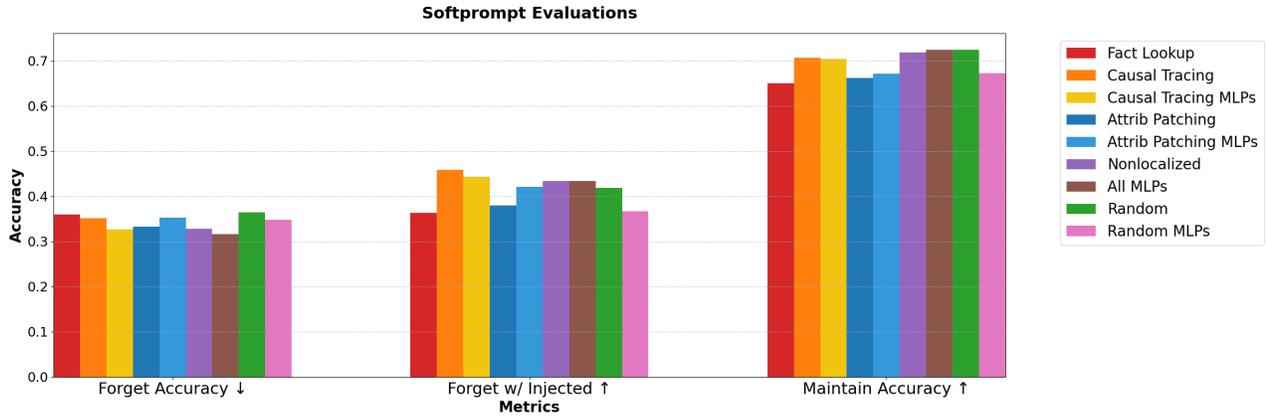


Figure 47. Metrics with soft prompts applied for CounterFact-Editing, averaged over all models.

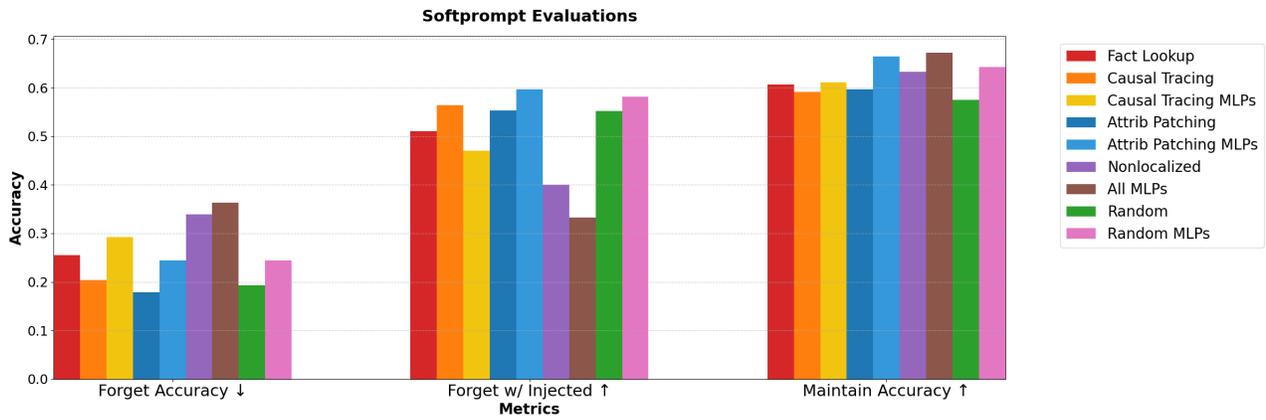


Figure 48. Metrics with soft prompts applied for Sequential-CounterFact-Editing, averaged over all models.