Robust Knowledge Unlearning via Mechanistic Localizations

Phillip Guo [*1] **Aaquib Syed** [*1] **Abhay Sheshadri** [2] **Aidan Ewart** [3] **Gintare Karolina Dziugaite** [4]

Abstract

Methods for machine unlearning in large language models seek to remove undesirable knowledge or capabilities without compromising general language modeling performance. This work investigates the use of mechanistic interpretability to improve the precision and effectiveness of unlearning. We demonstrate that localizing unlearning to components with particular mechanisms in factual recall leads to more robust unlearning across different input/output formats, relearning, and latent knowledge, and reduces unintended side effects compared to nonlocalized unlearning. Additionally, we analyze the strengths and weaknesses of different automated (rather than manual) interpretability methods for guiding unlearning, finding that their corresponding unlearned models require smaller edit sizes to achieve unlearning but are much less robust.

1. Introduction

Large language models (LLMs) often learn to encode undesirable knowledge, such as generating harmful stereotypes or leaking private information. The ability to selectively "unlearn" this problematic knowledge is paramount for ensuring safety, fairness, and control of AI. Yet, removal of knowledge from these models presents significant challenges.

Unlearning methods often rely on gradient-based updates of the entire model. These often come at the cost of affecting other general or tangential knowledge within the model. Moreover, the unlearning achieved through these methods may not be robust – slight variations in the prompt formulation can often still elicit the unlearned fact or capability.

Some recent editing and unlearning techniques use heuristics to judge the relevance of model components (such as specific neurons or MLP modules) for particular facts, and then localize updates to those components to minimize undesired side effects of unlearning. However, Hase et al. (2023) has cast doubt on the efficacy of these importance heuristics for edits and unlearning. Indeed, our work finds that unlearning of facts based on common *automated localization techniques*, even ones which are designed to find causally important components, performs no-better or even worse than non-localized unlearning, especially when evaluated on a broad-range of tests that check for robustness and latent information in intermediate representations.

In contrast to these automated localization techniques, mechanistic interpretability seeks to attribute describable task mechanisms to particular components. We hypothesize that this detailed understanding of component mechanisms, requiring what we refer to as *manual mechanistic interpretability*, leads to improved localization for unlearning.

In this work we exploit the findings from this literature to design a manual interpretability-based unlearning technique that we refer to as *manual mechanistic unlearning*. Using this, we localize our unlearning and cause the model to robustly unlearn facts with minimal side effects. Our experiments demonstrate that manual mechanistic unlearning outperforms all other automated localization and nonlocalized unlearning approaches.

Summary of Contributions

- We motivate the necessity for robust unlearning approaches and evaluation by demonstrating the inability of standard approaches to generalize to output distribution shifts, adversarial relearning, and latent knowledge probing on our chosen task.
- We demonstrate that localization informed by manual mechanistic interpretability leads to robust and targeted unlearning that generalizes well and is resilient to relearning and probing. In contrast, automated localization and nonlocalized unlearning approaches are not as robust/targeted.
- We show that both automated localization and manual mechanistic interpretability approaches can achieve a given forget set inaccuracy with smaller edit sizes (fewer weights masked).

[*] Equal contribution, determined by coin flip [1] University of Maryland [2] Georgia Institute of Technology [3] University of Bristol [4] Google DeepMind. Correspondence to: Phillip Guo <phguo@umd.edu>, Aaquib Syed <asyed04@umd.edu>, Gintare Karolina Dziugaite <gkdz@google.com>.

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 factual recall interpretability literature from (Nanda et al., 2023; Geva et al., 2023; Chughtai et al., 2024; Yu et al., 2023), which discovers mechanisms for the retrieval and formatted extraction of factual information.

Automated Circuit Discovery 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 find these subnetworks that are most critical for the model's behavior 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 Machine unlearning seeks to modify pre-trained models to eliminate or alter learned knowledge such as capabilities or facts. A number of prior approaches focused 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). This goal of unlearning to match a retrained-from-scratch model relates to a mathematical definition for unlearning that have been proposed by Ginart et al. (2019), closely relating to 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) identifies 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).

Evaluating Failures of Unlearning Several recent papers demonstrate failures of unlearning/editing methods, both localized and nonlocalized. 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.

2. Methods

2.1. Unlearning Tasks

We focus on unlearning subsets of the Sports Facts dataset from Nanda et al. (2023), which contains subject-sport relations across three sports categories for 1567 athletes.

We attempt to unlearn 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 unlearn a set of 16 athletes belonging to all three sport categories. The forget set here is the set of the 16 athletes. In both groups, the retain sets are the rest of the non-forget athletes.

For all tasks, 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.

2.2. Unlearning Procedure

For all unlearning, we use a two-step unlearning process. First, we use a localization method to isolate a subset of components that we deem are valuable or meaningful candidates for unlearning. Then, we optimize over the parameters of these isolated components using an unlearning loss function, with the end goal being a model that is incapable of performing the associations in the forget set while retaining other retain associations and general language modeling capability.

2.3. 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 $S : C \to \mathbb{R}$, 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 5% of the total components of the model, corresponding to approximately 10 components.

We use efficient automated localization methods for finding these mappings currently available in the literature.

Causal Tracing and Attribution Patching First, we test Causal Tracing, a method for finding components with high direct 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.

Manual Mechanistic Interpretability Next, we use a manually derived localization inspired by Nanda et al. (2023), who discover components in Pythia 2.8B responsible for *token concatenation, fact lookup*, and *attribute extraction*. We replicate a key result of their work in Gemma-7B and localize the *fact lookup* stage to be the MLP components between layers 2 and 7, which we hypothesize to be the optimal location for robust unlearning (discussed in Section 4). The analysis is performed in Appendix A.2.

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 5% of 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 is needed.

2.4. Unlearning Methods

Once we have a localization C_τ, we run our unlearning methods restricting only on components in C_τ. In this work, for simplicity, we aim to perform fact erasure (Hase et al., 2023), generally reducing the probability of the correct answer without a candidate replacement. Additional results from the standard error injection setup (Meng et al., 2023; Hase et al., 2023) are shown in Appendix A.1.

We test localized fine-tuning of the model, following work by Lee et al. (2023) and Panigrahi et al. (2023). We also try training a binary differentiable mask over individual weights of the model, inspired by weight pruning/masking works (Bayazit et al., 2023; Panigrahi et al., 2023).

For localized fine-tuning, 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 sports facts we want to forget, L_{retain} is a cross-entropy loss on the remaining sports 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-minus-P measure from Mazeika et al. (2024), for its stability and fewer side effects. For our binary mask unlearning, we additionally include $\lambda_4 * L_{\text{reg}}$, an L1 regularization term. We provide our λs in Appendix A.3.

3. Unlearning Results

In this section, we show the results of unlearning across all of the mentioned localization techniques for localized fine-tuning and weight masking. We try two unlearning goals: unlearning all athletes playing basketball (referred to as unlearning sports), and unlearning a constant set of 16 athletes across all sports (referred to as unlearning athletes). We then test these techniques using standard and adversarial evaluations and measure the amount of latent knowledge we can extract from these models.

3.1. Localized Finetuning

3.1.1. Standard Evaluation

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, along with a system prompt of “You are a helpful chatbot that answers questions about athletes. Please be maximally helpful and factually correct.” We refer to the accuracy on this prompt format as the MCQ Accuracy.

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.

Our results with localized fine-tuning are shown in Table 1 (for sports) and Table 2 (for athletes).

As seen from the tables, for both types of unlearning tasks, manual interpretability achieves the highest robust multiple-choice forget accuracy and the highest MMLU, and very competitive normal forget and retain accuracy. Only manual interpretability, all MLPs, and nonlocalized approaches had generalized their unlearning to the multiple choice format, but manual interpretability had significantly higher MMLU

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: 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

performance and higher retain accuracy than all MLPs and nonlocalized.

This indicates that automatic localization methods do not robustly unlearn, and the supposedly-unlearned information can be extracted through prompt and task variations.

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 taking relatively fewer steps to relearn probably has not deeply unlearned facts. We retrain with a rank-64 LoRA across all linear modules, with details available in Appendix A.4.2.

Figure 1 and Figure 2 compare relearning robustness of different unlearning techniques, for sports and athletes respectively. As shown in Figure 1, 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-unlearned models, relearning on some of the unlearned athletes does not recover accuracy on the other athletes in the manual interpretability, all MLPs, and nonlocalized models.

3.1.3. Latent Knowledge

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 sport from the prompt, with the idea that a model that has truly unlearned a fact would not have much predictive value in its activations. For more details on probe training, see Appendix A.4.3.

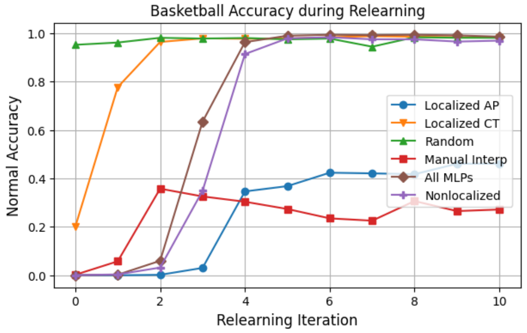

Figure 1. Retraining basketball-unlearned models with two athletes in the forget set, for ten iterations.

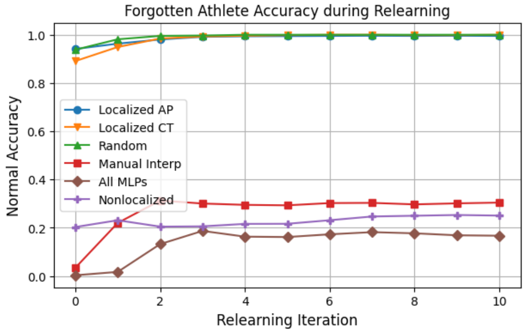

Figure 2. Retraining athlete-unlearned models with two athletes in the forget set, for ten iterations. The y-axis represents the normal accuracy on the forget set. Low-resource relearning of athletes demonstrates that manual and non-localized unlearning techniques are robust to this test (staying close to the guessing rate of 33%), while the other methods maintain full performance on the entire forget set.

We test whether the models post-unlearning retain information about the forget set in the intermediate layer representations. Figure 3 and Figure 4 show the accuracy of the trained per-layer probes on the forget set for each of the unlearning tasks, sports and athletes respectively.

Figure 4 provides evidence that the manual interpretability, all MLPs, and nonlocalized athlete-unlearned models contain less or no recoverable representations of the unlearned association in the intermediate layers.

Figure 3 shows that all unlearning techniques we test on basketball produce models that continue to encode the supposedly-unlearned basketball associations even while these models output an incorrect answer (Tables 1 and 2). However, this is somewhat expected because unlearned models likely learn to treat basketball prompts distinctly from

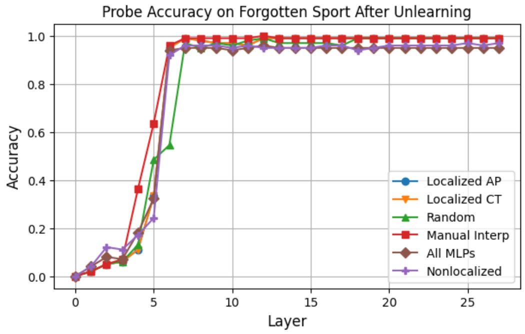

Figure 3. **Probe accuracy (combined over all three sports) on the basketball-unlearned models, by layer.** Probing reveals that all unlearning methods leave recoverable information about which sport is the answer, which is expected because the unlearned models likely treat all basketball prompts significantly differently from the retained sport prompts. The probe accuracies for Attribution Patching, Causal Tracing, and Random localized-models overlap because their different unfrozen components are all in later layers, when the probe accuracies are already 100%.

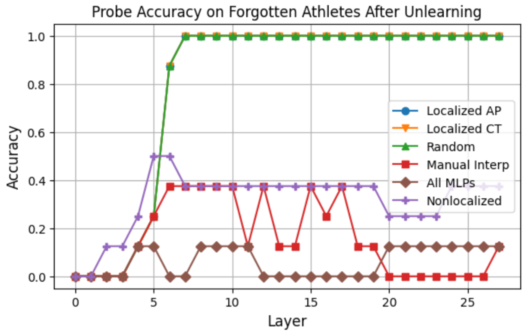

Figure 4. Probe accuracy (combined over all three sports) on the athlete-forgotten models across layers.

non-basketball prompts, which makes it no less difficult for probes to learn to distinguish these prompts.

Note, however, that due to the nature of this experiment and dependence on the probing technique used, a positive result (failing to reconstruct the accuracy based on intermediate layer representations) does not rule out that the forget set is still represented in these layers, and that a different probing/extraction technique may be more successful. A negative result, on the other hand, is conclusive: if the probe accuracy is high, we know the unlearned information is still recoverable in the intermediate layer representations.

3.2. Weight Masking

The localization methods discussed above isolate different numbers of components/parameter counts, resulting in unlearning techniques that may vary in terms of the total number of modifiable parameters. In this section we employ weight masking to quantify the size of edits needed to unlearn facts, for more direct comparisons. In particular, we empirically evaluate how a learned binary mask over the individual weights of the localized components can produce unlearning, and vary the size of this mask/number of masked elements.

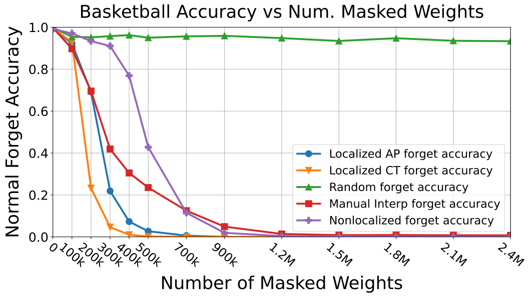

Figure 5. Testing the models' unlearning of basketball athletes against the number of weights masked.

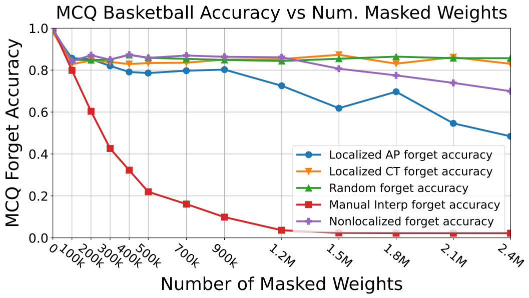

Figure 6. Testing the models' unlearning of basketball athletes against the number of weights masked, in the MCQ prompt format.

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. In particular, Attribution Patching and Causal Tracing localizations cause the model to have zero accuracy on the in-distribution set with much fewer masked weights needed than every other localization, including manual interpretability.

However, 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 6).

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.

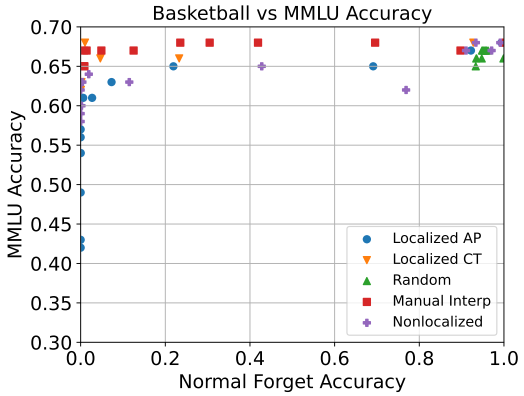

Figure 7. Unlearning basketball facts, measuring MMLU and forget set performance across different discretization thresholds.

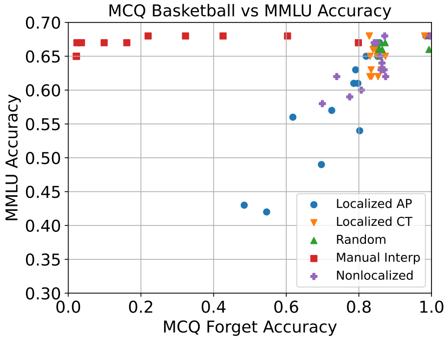

Figure 8. Unlearning basketball facts, measuring MMLU and MCQ forget set performance across different discretization thresholds.

(2024)). While all localized methods perform well when evaluated normally (Figure 7), Figure 8 shows manual localization generalizes for minimizing loss of MMLU capabilities while unlearning sports facts in the MCQ format, while other methods experience relatively significant side effects across different numbers of weights masked.

For unlearning the subset of athletes, Figure 9 shows that causal tracing localization causes the model to have 0% accuracy on the forget set, and manual interpretability and nonlocalized unlearning cause the model to have near guessing rate (33%) accuracy. However, only manual localization minimizes loss of capabilities while unlearning the athlete subset (Figure 11).

Furthermore, no method completely generalizes this unlearning to the MCQ prompt format (Figure 10), and manual localization remains superior in minimizing loss of capabilities while unlearning the athlete subset (Figure 12).

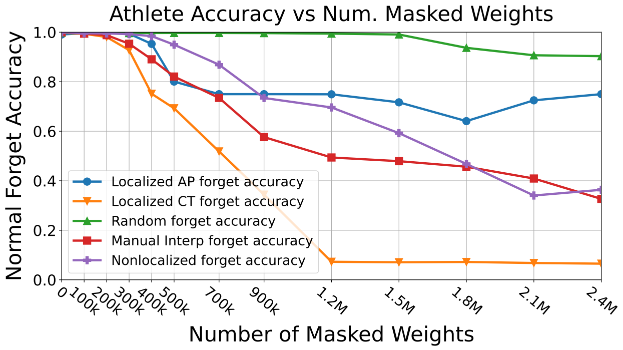

Figure 9. Unlearning subset of athletes, measuring accuracy on the forget set.

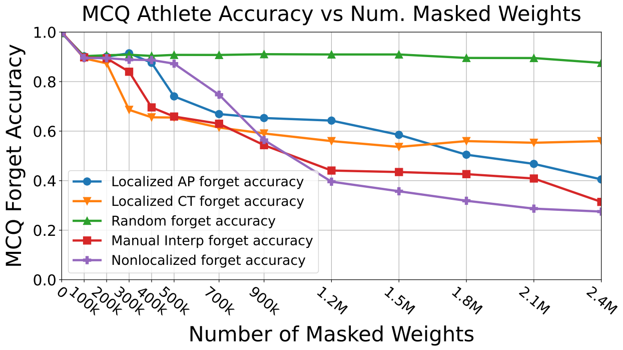

Figure 10. Unlearning subset of athletes, measuring accuracy on the forget set in the MCQ prompt format.

4. Discussion

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

Our work evaluates the efficacy of different localization methods for unlearning 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.

We hypothesize that automated localization 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 re-

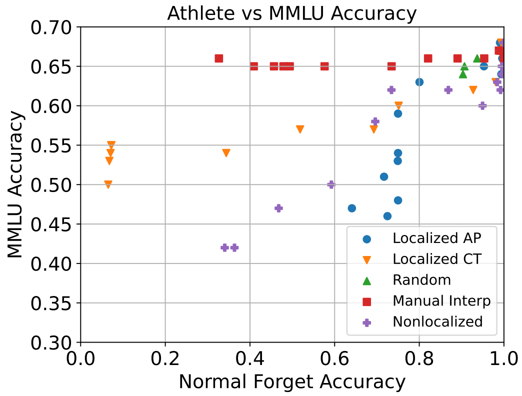

Figure 11. Unlearning subset of athletes, measuring MMLU and forget set performance across different discretization thresholds.

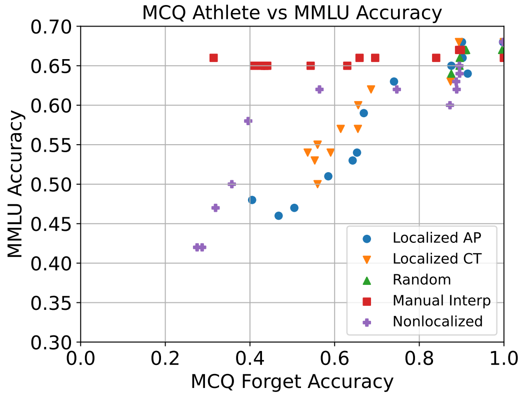

Figure 12. Unlearning subset of athletes, measuring MMLU and MCQ forget set performance across different discretization thresholds.

pair the original attention mechanism. In contrast, mechanistic understanding allows us to target unlearning at the sites where knowledge is sourced, which we hypothesize to robustly prevent that information from entering the latent stream in any format.

Our models generalize across prompt formats and resist adversarial relearning and latent probing while minimizing side effects.

Our work also suggests unlearning 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 the quality of localizations and explanations.

Impact Statement

This paper advances the fields of interpretability and unlearning, both of which are relevant for ensuring the safety, privacy, and fairness of models. We hope our methods help model developers responsibly unlearn 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.

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.

Chen, J. and Yang, D. Unlearn what you want to forget: Efficient unlearning for llms. *arXiv preprint arXiv:2310.20150*, 2023.

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.

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.

Eldan, R. and Russinovich, M. Who's harry potter? approximate unlearning in llms. *arXiv preprint arXiv:2310.02238*, 2023.

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.

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.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization, 2017.

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.

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, 2024.

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.

Mallen, A. and Belrose, N. Eliciting latent knowledge from quirky language models, 2023.

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., Kramar, 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.

Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., and Henighan, T. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

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.

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 3 and Table 4, 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 3. 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 4. Results of unlearning 16 athlete associations, with the objective of replacing the correct sport with ”Golf”.

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

We also perform relearning and latent knowledge experiments in Figure 13 and Figure 14, demonstrating that manual localization for the athlete subset injection improves relearning and latent knowledge robustness.

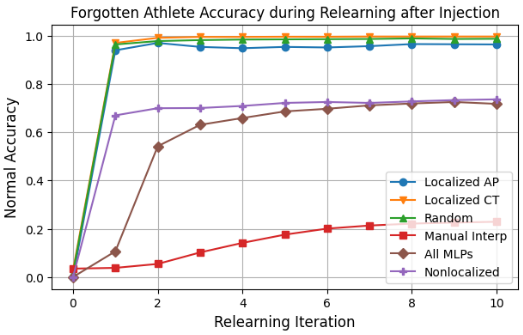

Figure 13. Retraining models that have had 16 athlete-sport associations replaced with Golf as the athletes’ associated sport, with two athletes in the forget set, for ten iterations.

A.2. Gemma Interpretability Analysis

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 16).

Unlike Nanda et al. (2023), we find attention heads past layer 2 that impact the linear representation of attributes and thus could potentially be important for fact lookup (Figure 15). However, because they could likely play a variety of other different roles, 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.

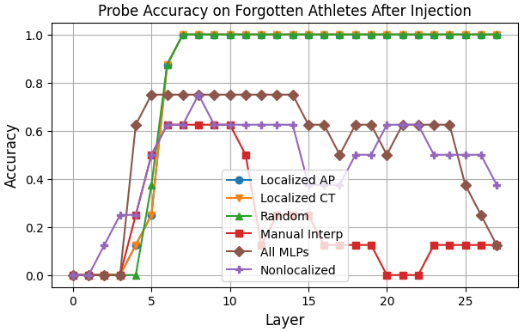

Figure 14. Probe accuracy (combined over all three sports) on the athlete golf-injected models across layers.

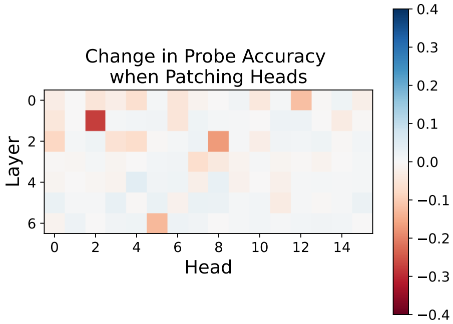

Figure 15. Difference in final layer probe accuracy when mean ablating a single head for all heads between layers 0 and 6.

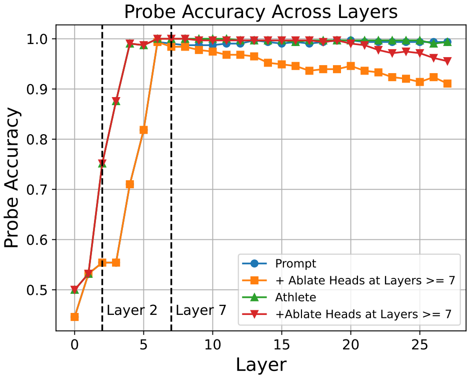

Figure 16. Probe accuracy on predicting sport across layers. ”Prompt” refers to the entire facts prompt, while ”Athlete” is just the athlete’s name.

A.3. 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, 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 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 $1 * 10^{-3}$ for both basketball and athletes and clamping the mask values between 0 and 1 every update step.

A.4. Evaluation Details

A.4.1. 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.4.2. 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.4.3. 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.