BIDIRECTIONAL CAUSAL TRACING AND INTERVENTION: CAUSAL MECHANISM-GUIDED MACHINE UNLEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model unlearning is commonly evaluated through changes in answer probabilities or output behavior, but such behavioral suppression does not demonstrate that the causal support underlying target knowledge recall has been weakened. Reliable unlearning therefore raises two coupled questions: where is this causal support located, and how should it be weakened? To address the first, we use causal tracing under forward and reverse queries to localize internal components whose restorative effects persist across both retrieval directions. To address the second, we optimize localized adapters to directly reduce these components' causal restoration effects, while directional balancing and retention constraints limit collateral damage to non-target knowledge. We call this framework Bidirectional Causal Tracing and Intervention (BCTI). On 50 target facts with Pythia-6.9B-deduped, BCTI achieves an average 87.8% reduction in target answer scores, while a held-out causal audit records a $98.6\% \pm 1.5\%$ reduction in causal restoration effect. Applying the same procedure to OLMo-7B yields a $97.8\% \pm 1.9\%$ causal restoration reduction. These results support weakening and auditing causal recall support as a more reliable basis for assessing forgetting than behavioral suppression alone.

1 INTRODUCTION

The ability of Large Language Models (LLMs) to internalize and reproduce training knowledge supports broad generation capabilities, but can also expose private data, copyrighted content, and harmful knowledge, raising privacy, security, and compliance concerns (Yao et al., 2024b). Because retraining for every knowledge deletion request is prohibitively expensive, LLM unlearning aims to remove the influence of target knowledge through post-training intervention while preserving non-target capabilities (Dorna et al., 2025; Wang et al., 2025). Existing evaluations mainly rely on answer probabilities, model utility, and privacy attacks, which do not directly measure whether the internal mechanisms supporting target recall have weakened (Dorna et al., 2025). This distinction is consistent with causal representation learning, which separates predictive associations from their underlying mechanisms (Schölkopf et al., 2021). Unlearning should therefore be evaluated through both behavioral changes and changes in the causal support for target recall.

This distinction also helps organize existing LLM unlearning methods. The first category directly suppresses forget-set outputs through gradient ascent, alternative answers, or negative preference objectives (Wang et al., 2025; Zhang et al., 2024; Liu et al., 2024). These objectives reduce target-answer probabilities without directly measuring changes to internal recall mechanisms, and may create conflicts between forgetting and retention (Wang et al., 2025; Zhang et al., 2024). The second category uses causal or mechanistic analysis to guide interventions on recall-related components (Wang et al., 2027; Guo et al., 2025). In causal-tracing-based approaches, input corruption and activation restoration test which components recover the target prediction, providing a causally grounded localization criterion (Meng et al., 2022; Wang et al., 2027).

Our central claim is that disrupting the causal support underlying target knowledge recall supports a more reliable assessment of forgetting than behavioral suppression alone. Apparently forgotten knowledge can remain accessible through alternative prompts or benign relearning (Hu et al., 2025;

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

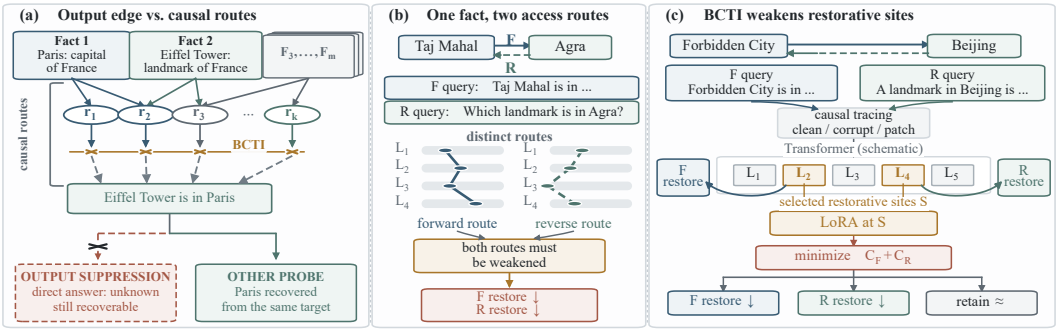


Figure 1: BCTI motivation. (a) Output suppression can preserve causal recall support. (b) Forward and reverse access require separate measurements. (c) BCTI localizes and weakens bidirectional restoration under retention constraints. All panels are conceptual.

Sun et al., 2025). Figure 1(a) illustrates why reduced answer accessibility should therefore be accompanied by weakened internal causal support.

This requirement must also account for retrieval direction. As illustrated by the Taj Mahal–Agra example in Figure 1(b), forward and reverse queries provide distinct views of the same association (Berglund et al., 2024; Xu et al., 2026). Weakening recall in one direction does not establish the same effect in the other, motivating causal tracing and evaluation in both directions. Yet tracing both directions only identifies causal support; it does not weaken that support or automatically determine an effective intervention (Meng et al., 2022; Hase et al., 2023). Existing causal and mechanistic approaches apply conventional forgetting objectives to selected components (Wang et al., 2027; Guo et al., 2025). We instead use causal restoration to specify both where to intervene and what to reduce.

We propose **Bidirectional Causal Tracing and Intervention (BCTI)** to implement this principle. Following Figure 1(c), BCTI selects sites with stable restorative effects in both directions and directly reduces these effects through localized adapters (Hu et al., 2022). Calibration balances directional signals, while reference-model KL divergence and gradient constraints limit changes to retained knowledge (Yu et al., 2020). To test whether causal weakening extends beyond the optimization prompts, BCTI separates fit, validation, and test prompts before localization. These splits respectively support localization and training, checkpoint selection, and held-out evaluation of behavioral forgetting, bidirectional causal restoration, and retention.

Under this protocol, BCTI achieves an 87.8% target-answer reduction and a $98.6\% \pm 1.5\%$ causal restoration reduction, with retain KL 0.006 (Tables 1 and 2). The causal-only ablation still achieves reductions of 83.4% and 97.9%, respectively, linking causal weakening to behavioral forgetting without an answer-level objective. Additional evaluation on OLMo-7B reproduces the strong causal restoration reduction, indicating that the central mechanistic finding is not specific to the Pythia parameterization. Comparisons with NPO and directional ablations further support the measured forgetting–retention trade-off and bidirectional design.

The main contributions of this paper are as follows:

1. We distinguish behavioral suppression from weakened causal support for factual recall as a basis for more reliable evaluation of unlearning.
2. We propose bidirectional causal tracing to jointly select intervention sites while preserving direction-specific measurements for optimization and evaluation.
3. We develop BCTI, which integrates causal restoration optimization, localized intervention, retention constraints, and held-out verification into a complete unlearning pipeline.

2 Related Work

Large Language Model Unlearning. Existing LLM unlearning methods primarily reduce access to target knowledge while preserving non-target capabilities. Common approaches include gradient ascent on forget samples, negative preference optimization, joint optimization over forget and

retain sets, and more localized updates based on parameter importance, gradient information, or knowledge-related subspaces (Yao et al., 2024a; Zhang et al., 2024; Fan et al., 2025; Tian et al., 2024; Lizzo & Heck, 2025). Although these methods differ in how they constrain model updates, they are generally evaluated through changes in output probabilities and downstream behavior.

However, behavioral suppression alone does not establish reliable forgetting. Forgotten information may remain detectable through privacy auditing, recover under benign relearning, or reappear under targeted probing (Du et al., 2025; Hu et al., 2025; Sun et al., 2025). Conversely, interventions localized to components involved in factual recall have been shown to produce more robust forgetting and stronger resistance to relearning (Guo et al., 2025). Together, these findings motivate evaluating unlearning not only by whether the target answer disappears, but also by whether the causal support for factual recall is weakened.

Causal Tracing and Mechanistic Localization. Causal tracing is a mechanistic localization technique for estimating the contribution of internal model components to a specific prediction. It corrupts the computation associated with a target input to suppress the original prediction and then restores clean activations at selected layers or token positions. The extent to which the target prediction is recovered is evidence of the restored component's causal contribution (Meng et al., 2022).

Causal tracing and related activation-patching methods have been widely used to study factual recall, trace information flow through transformer layers, identify layers or modules involved in particular predictions, and select intervention sites for model editing. Subsequent work has also systematized methodological choices such as corruption strategies, patching locations, and evaluation metrics, establishing activation patching as a general tool for mechanistic localization and circuit analysis (Zhang & Nanda, 2024).

Existing causal-tracing analyses estimate causal restoration effects within a specific corrupted input-output computation (Meng et al., 2022; Zhang & Nanda, 2024). Consequently, the resulting causal profile characterizes the internal computations supporting that particular prediction rather than a direction-independent representation of the underlying fact. Prior studies on the reversal curse show that successful retrieval from subject to object does not guarantee retrieval from object to subject (Berglund et al., 2024), while bidirectional editing studies report similar inconsistencies after factual modification (Xu et al., 2026). These findings suggest that forward and reverse queries should be treated as distinct mechanistic views of a factual association.

Causal Localization for Machine Unlearning. Mechanistic and causal perspectives have recently been introduced into machine unlearning to better identify and control the internal sources of target knowledge. Liu et al. (2024) formulate targeted unlearning through a causal-intervention framework, while EXACT uses causal tracing to attribute target knowledge to specific experts in mixture-of-experts language models and restrict updates to these components (Wang et al., 2027).

In the approaches summarized above, causal analysis formulates the intervention or determines where parameter updates should be applied, whereas the subsequent forgetting process remains driven by conventional objectives over model outputs, losses, or localized parameter changes. Consequently, causal analysis primarily determines where to intervene, while the causal contribution itself is not directly optimized as the forgetting target. The causal signal therefore mainly serves as a localization criterion rather than the primary quantity whose reduction defines the intervention.

In contrast, our method directly treats the causal contribution underlying factual recall as the object of intervention. We characterize each target relation through forward and reverse causal tracing and optimize the model to reduce the resulting bidirectional causal restoration scores. Answer-level suppression and retention regularization are incorporated only as auxiliary constraints, so behavioral suppression is not treated as sufficient evidence of reliable forgetting.

3 Preliminaries and Problem Formulation

Targeted Factual Knowledge Unlearning Given a pretrained language model f_θ, a forget set $\mathcal{D}_f$ containing target factual associations, and a retain set $\mathcal{D}_r$ containing unrelated knowledge, targeted factual knowledge unlearning aims to obtain an updated model f_{θ^-} that weakens recall of the facts in $\mathcal{D}_f$ while preserving the model's behavior on $\mathcal{D}_r$.

We focus on post-training intervention for factual knowledge rather than exact deletion of individual pretraining examples or full retraining without the forget data. Accordingly, the goal is to suppress the influence of specified factual associations with minimal degradation of unrelated model capabilities, without claiming strict equivalence to retraining on the retained data alone.

Bidirectional Factual Queries Each factual association is evaluated in two directions. Given a subject–relation–object triple (s, r, o), the forward query $q^{\rightarrow}$ takes the subject s and relation r as input and predicts the corresponding object o, whereas the reverse query $q^{\leftarrow}$ starts from the object o and retrieves the associated subject s under the inverse relation r^{-1}. Formally,

$$q^{\rightarrow}(s, r) \rightarrow o, \qquad q^{\leftarrow}(o, r^{-1}) \rightarrow s.$$

The two directions are treated as distinct query formulations of the same factual association because successful retrieval under one direction does not guarantee that the inverse association is represented or accessed in the same manner (Berglund et al., 2024; Xu et al., 2026).

Causal Tracing for Factual Recall Causal tracing estimates the contribution of an internal model component to factual recall by comparing clean, corrupted, and activation-patched forward passes (Meng et al., 2022). For a target answer sequence $y = (y_1, \ldots, y_T)$, we define a length-normalized target-answer score using the geometric mean of the conditional token probabilities:

$$p(y \mid x) = \exp\left(\frac{1}{T}\sum\nolimits_{t=1}^{T} \log p(y_t \mid x, y_{<t})\right).$$

Let $p_{\text{clean}} = p(y \mid x)$ and $p_{\text{corrupt}} = p(y \mid \tilde{x})$ denote the target-answer scores under the original prompt x and the corrupted prompt $\tilde{x}$, respectively. For an intervention at layer l and position or module s, $p_{\text{patch}}^{(l,s)}$ is the target-answer probability obtained by restoring the corresponding clean activation within the corrupted forward pass.

The normalized causal restoration score is defined as

$$C^{(l,s)} = \begin{cases} \text{clip}\left(\dfrac{p_{\text{patch}}^{(l,s)} - p_{\text{corrupt}}}{p_{\text{clean}} - p_{\text{corrupt}}}, 0, 1\right), & p_{\text{clean}} - p_{\text{corrupt}} > \varepsilon, \\ 0, & p_{\text{clean}} - p_{\text{corrupt}} \leq \varepsilon \end{cases},$$

where ε is a small constant used to exclude cases in which corruption does not produce a meaningful probability decrease. A score of zero indicates that restoring the selected activation does not recover the target-answer probability removed by corruption, whereas a score of one indicates recovery to at least the clean-prompt level.

We use the same causal restoration score definition during localization, optimization, and pre- and post-unlearning auditing. During training, p_{clean} and p_{corrupt} are treated as stop-gradient baselines, so gradients are propagated only through the activation-patched computation.

4 METHOD

4.1 METHOD OVERVIEW

Behavioral suppression alone cannot determine whether unlearning weakens factual recall or merely lowers the target answer probability. BCTI therefore targets the causal support of recall through three stages (Figure 2): bidirectional causal localization determines where intervention is permitted; causal-score-guided intervention specifies what internal capability must weaken; and balanced, retention-constrained optimization controls how this reduction is achieved.

This ordering is necessary: localization alone does not weaken recall, one-directional intervention may leave an alternative access route intact, and unconstrained updates may disrupt shared computations (Wang et al., 2025; Yu et al., 2020).

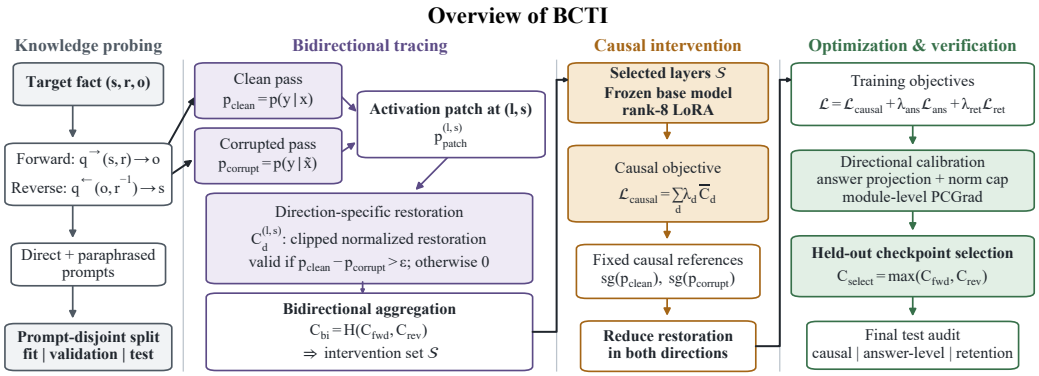


Figure 2: BCTI pipeline: prompt-disjoint bidirectional queries, causal localization, selected LoRA intervention, balanced optimization, and held-out verification.

4.2 Bidirectional Causal Localization

Activation strength or association does not establish causal involvement (Vig et al., 2020). BCTI instead uses activation patching to identify components whose restoration recovers the target answer after input corruption (Meng et al., 2022; Zhang & Nanda, 2024). For each fact, forward and reverse queries are traced separately because the same association may rely on directionally asymmetric retrieval pathways (Berglund et al., 2024; Xu et al., 2026).

Figure 3(a) makes this distinction concrete. A forward prompt can activate a generic country–currency completion process that favors currency-type continuations without uniquely encoding the Japan–yen binding. Weakening such a component may suppress *yen* after *Japan* while leaving the reverse query answerable. The reverse prompt instead tests whether *yen* can retrieve *Japan*. Requiring restoration in both directions therefore demotes one-way completion features and favors components that participate in retrieving the same association from either endpoint, consistent with progressive subject–relation integration (Geva et al., 2023).

Because patching estimates can vary across prompts and corruption instances (Zhang & Nanda, 2024; Canby et al., 2025), scores are robustly aggregated within each direction. Let $C^{(l,s)}_{\text{fwd}}$ and $C^{(l,s)}_{\text{rev}}$ denote the resulting scores for component (l, s). We combine them using

$$C^{(l,s)}_{\text{bi}} = \begin{cases} \dfrac{2C^{(l,s)}_{\text{fwd}} C^{(l,s)}_{\text{rev}}}{C^{(l,s)}_{\text{fwd}} + C^{(l,s)}_{\text{rev}} + \varepsilon}, & C^{(l,s)}_{\text{fwd}} > 0 \quad \text{and} \quad C^{(l,s)}_{\text{rev}} > 0, \\ 0, & \text{otherwise} \end{cases}.$$

The harmonic mean implements this distinction: a component is ranked highly only when its restorative role persists across both access directions. Stable high-scoring components define the intervention set $\mathcal{S}$, preventing a generic or unusually strong pathway in one direction from determining where the model is updated.

4.3 Causal-Score-Guided Intervention

Localization identifies states that can restore the answer but does not weaken that ability. BCTI therefore minimizes their direction-specific restoration scores:

$$\mathcal{L}_{\text{causal}} = \sum\nolimits_{d \in \{\text{fwd,rev}\}} \lambda_d \frac{1}{|\mathcal{S}|} \sum\nolimits_{(l,s) \in \mathcal{S}} C^{(l,s)}_d,$$

where fit-split calibration sets λ_d to balance the two directions. The harmonic mean is used only for localization: minimizing it directly could reduce one route while leaving the other active, whereas the separate terms require both restorative pathways to weaken.

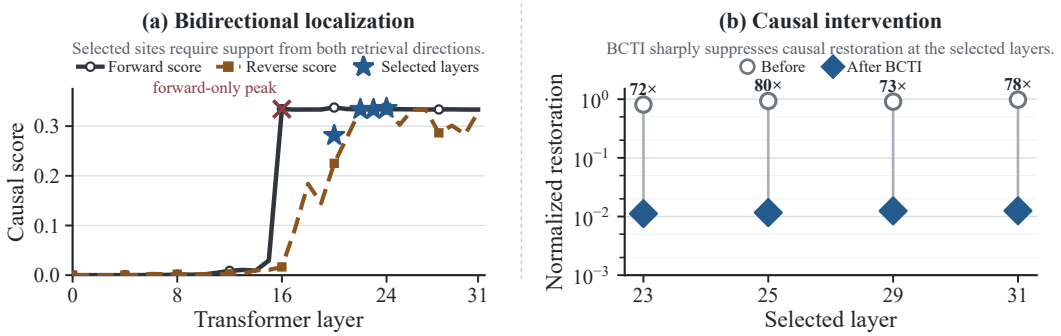


Figure 3: BCTI toy example. (a) Bidirectional localization excludes a forward-only peak. (b) Intervention reduces restoration at selected layers.

Figure 3(b) illustrates what this objective changes. Corrupting the subject removes the *France* clue; patching a selected state then tests whether the restored information can propagate through the remaining computation and rebuild the France–Paris association (Geva et al., 2023). The selected state is therefore defined by its restorative role rather than assumed to be an isolated storage location. An answer-only loss could make *Paris* less likely under the exact prompt while preserving this restorative ability; minimizing $C_d^{(l,s)}$ instead directly weakens the patched state's capacity to reconstruct the target answer. Because the clean and corrupted probabilities are stop-gradient references, the score is reduced through the activation-patched computation rather than through a moving normalization baseline. The base model remains frozen, with LoRA adapters inserted only into modules associated with $\mathcal{S}$.

4.4 Balanced Optimization and Causal Verification

Directional asymmetry and shared components can produce imbalanced forgetting and conflicts with retained knowledge (Berglund et al., 2024; Xu et al., 2026; Wang et al., 2025). BCTI therefore uses

$$\mathcal{L} = \mathcal{L}_{\text{causal}} + \lambda_{\text{ans}}\mathcal{L}_{\text{ans}} + \lambda_{\text{ret}}\mathcal{L}_{\text{ret}},$$

where the causal term is primary, the unlikelihood-style answer term suppresses the target sequence (Li et al., 2020), and the retention term constrains KL divergence from the frozen reference model on retain prompts.

The three terms define the optimization signals, but their gradients are composed hierarchically rather than treated as an unconstrained weighted sum. Any answer-gradient component conflicting with the causal gradient is projected out, and the remainder is norm-capped relative to the causal gradient. The resulting forgetting gradient is then combined with the retention gradient using module-level PCGrad (Yu et al., 2020). This preserves causal reduction as the internal target while using answer suppression and retention as behavioral constraints.

Finally, prompts are divided before localization into disjoint fit, validation, and test splits. The fit split supports localization, directional calibration, and optimization; validation selects checkpoints by the larger remaining directional restoration score; and the isolated test split evaluates held-out forward, reverse, and paraphrased forgetting together with causal restoration and retention.

5 Experiments

5.1 Experimental Setup

Models, data, and baselines. We evaluate BCTI on Pythia-6.9B-deduped (Biderman et al., 2023), independently unlearning 50 target facts spanning capital, location, and currency relations while using 100 disjoint retain facts. Forward, reverse, and paraphrased prompts follow the prompt-disjoint fit, validation, and test protocol in Section 4.4. BCTI is also evaluated on OLMo-7B (Appendix C).

Table 1: Comparison with six controlled baselines and BCTI over the same 50 forget targets. Values are mean ± sample standard deviation.

Source	Method	Forget (%) ↑	Retain KL ↓	Retain max (%) ↓	Updated parameters
[Jang et al., ACL 2023]	GA	100.0 ± 0.1	2.427 ± 1.689	89.9 ± 8.7	6.857B
[Yao et al., ACL 2024]	GA+KL	100.0 ± 0.1	0.655 ± 0.395	77.0 ± 17.2	6.857B
[Liu et al., CoLLAs 2022]	GradDiff	99.8 ± 0.8	1.137 ± 0.446	50.6 ± 23.8	6.857B
[Zhang et al., COLM 2024]	NPO	$\mathbf{100.0 \pm 0.0}$	1.310 ± 0.539	87.3 ± 11.1	6.857B
[Fan et al., NeurIPS 2025]	SimNPO	83.7 ± 21.7	1.744 ± 0.712	$\mathbf{10.5 \pm 13.2}$	6.857B
[jin et al., NeurIPS 2024]	NPO+LoRA	60.1 ± 34.1	0.025 ± 0.028	18.2 ± 16.3	**2.10M**
Ours	BCTI	87.8 ± 7.8	$\mathbf{0.006 \pm 0.004}$	12.5 ± 10.7	**2.10M**

We compare six controlled baselines: gradient ascent (GA) (Jang et al., 2023), GA with retain KL (GA+KL) (Yao et al., 2024a), gradient difference (GradDiff) (Liu et al., 2022), negative preference optimization (NPO) (Zhang et al., 2024), simplified NPO (SimNPO) (Fan et al., 2025), and rank-8 NPO+LoRA (jin et al., 2024). The baselines and BCTI share facts, splits, and prompts across all 50 targets; we therefore interpret the results as a matched controlled trade-off comparison.

Implementation and metrics. BCTI places rank-8 LoRA adapters in causally selected attention and MLP projections (Hu et al., 2022) and trains them with AdamW for at most 80 steps at 5×10^{-5} (Loshchilov & Hutter, 2019). We report target-answer reduction (Forget), the weaker direction (Min-dir.), causal-restoration reduction, retain KL, maximum retain-answer drop, and updated parameters.

These metrics test distinct parts of the claim. Forget measures behavioral suppression, whereas Min-dir. prevents strong performance in one direction from masking residual access in the other, a concern motivated by directional retrieval asymmetry (Berglund et al., 2024; Xu et al., 2026). Causal-restoration reduction audits whether patched states retain their previously measured contribution (Meng et al., 2022; Zhang & Nanda, 2024). Retain KL and maximum answer drop respectively capture average distributional drift and the worst observed factual degradation, which need not agree when forgetting and retention gradients conflict (Wang et al., 2025).

5.2 Main Results

Table 1 reports $(87.8 \pm 7.8)\%$ forgetting for BCTI over 50 targets. Among the six baselines evaluated over the same 50 targets, full-parameter NPO has the highest mean forgetting at $(100.0 \pm 0.0)\%$, but incurs retain KL 1.310 and an 87.3% maximum retain drop; BCTI records 0.006 and 12.5%, respectively, while updating 2.10M rather than 6.857B parameters. This matched comparison establishes the observed forgetting–retention and parameter-efficiency trade-off.

On held-out prompts, the causal audit shows a $(98.6 \pm 1.5)\%$ restoration reduction (Table 2), showing that answer suppression accompanies weakened restorative capacity. Figure 4 visualizes this comparison across forgetting, average retention drift, and worst-case retain damage.

5.3 Ablation Studies

Table 2 reports the exact objective, localization, and locator–objective results, while Figure 5 visualizes their forgetting–retention trade-offs. Answer-only optimization yields 94.7% forgetting, 98.6% causal reduction, and a 15.4% retain maximum drop; causal-only optimization yields 83.4%, 97.9%, and 11.7%, respectively. Full BCTI reaches 87.8%, 98.6%, and 12.5%. Thus, the causal objective alone substantially disrupts the audited pathway, whereas the hierarchical combination balances behavioral forgetting, causal reduction, and retention rather than maximizing a single endpoint.

Forward- and reverse-only localization reduce Min-dir. forgetting to 82.9% and 81.5%, versus 83.3% for BCTI, with causal reductions of 98.3% and 98.1% versus 98.6%. These modest gaps are consistent with directional complementarity (Section 4.2). The locator–objective substitutions are more decisive: ROME localization with the BCTI objective yields only 0.6% forgetting and

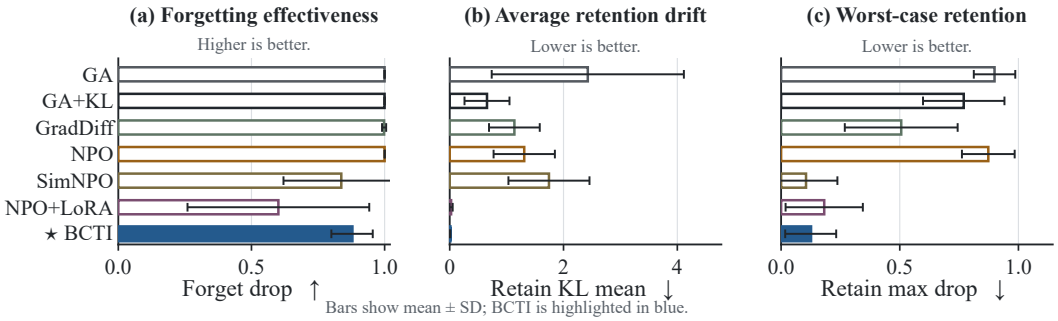


Figure 4: Forgetting–retention trade-off: (a) Forget, (b) retain KL, and (c) maximum retain-answer drop. Error bars show sample standard deviation.

Table 2: Ablation results over 50 forget targets (mean ± sample standard deviation).

Variant	Forget (%) ↑	Min-dir. (%) ↑	Causal restoration (%) ↑	Retain max (%) ↓
Answer-only obj.	94.7 ± 7.0	92.2 ± 9.3	$\mathbf{98.6 \pm 2.3}$	15.4 ± 15.1
Causal-only obj.	83.4 ± 9.1	78.1 ± 11.6	97.9 ± 1.9	11.7 ± 9.9
Forward-only loc.	87.2 ± 8.3	82.9 ± 10.6	98.3 ± 2.2	12.3 ± 9.9
Reverse-only loc.	86.1 ± 9.8	81.5 ± 12.2	98.1 ± 2.2	12.4 ± 8.6
ROME loc. + BCTI obj.	0.6 ± 4.3	-1.5 ± 4.8	2.1 ± 5.7	$\mathbf{11.7 \pm 5.1}$
BCTI loc. + PaLU obj.	$\mathbf{98.1 \pm 4.3}$	$\mathbf{97.0 \pm 6.3}$	91.1 ± 10.5	59.5 ± 29.1
BCTI	87.8 ± 7.8	83.3 ± 10.4	98.6 ± 1.5	12.5 ± 10.7

2.1% causal reduction (Meng et al., 2022), whereas the BCTI locator with PaLU yields 98.1% forgetting but only 91.1% causal reduction and 59.5% retain damage (Zhai et al., 2026). Thus, the observed trade-off depends jointly on where and how the model is updated.

5.4 Analysis of BCTI

Across the main results and ablations, strong behavioral forgetting is accompanied by sharply reduced ability of patched states to recover the answer. Bidirectional traces further reveal shared and direction-specific support, consistent with known retrieval asymmetries (Berglund et al., 2024; Xu et al., 2026). The lower Min-dir. results under one-direction localization motivate treating retrieval direction as part of the mechanism rather than only an evaluation variant.

Finally, the locator–objective substitutions distinguish three quantities that a single output score conflates: causal involvement, causal reducibility, and retention cost. Together, the evidence supports coupling bidirectional localization, causal optimization, and retention constraints, while the dispersion across targets cautions against inferring reliable forgetting from an aggregate behavioral score alone.

6 Discussion and Limitations

BCTI suggests that unlearning evaluation should distinguish behavioral inaccessibility from mechanistic disruption. Output-based tests alone may overestimate forgetting because apparently forgotten knowledge can remain accessible through targeted probing or rapid relearning (Hu et al., 2025; Sun et al., 2025). Causal restoration provides a complementary audit by testing whether internal states that previously supported recall remain capable of recovering the target answer (Meng et al., 2022; Zhang & Nanda, 2024). The across-target dispersion in Table 1 and the directional ablations in Table 2 show that forgetting outcomes vary across both facts and access directions. Because prior work also documents conflicts between forgetting and retention gradients (Wang et al., 2025), these two observations jointly motivate evaluating forgetting, retention, and causal restoration reduction rather than inferring success from any one quantity.

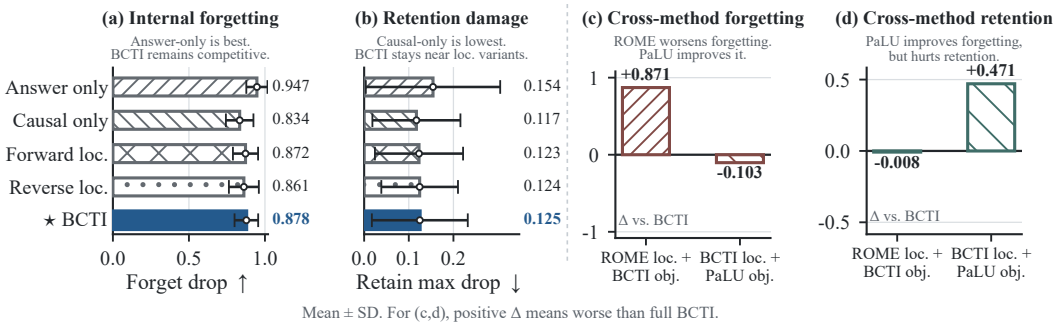


Figure 5: BCTI ablations: (a,b) internal variants and (c,d) locator–objective substitutions relative to full BCTI. Positive Δ denotes worse performance; error bars show sample standard deviation.

However, a reduced causal restoration score does not prove that all representations of the target fact have been erased. Causal tracing measures restoration within a particular corrupted computation, and its conclusions depend on the prompts, corruption strategy, patched components, and scoring function (Meng et al., 2022; Zhang & Nanda, 2024). Distributed, redundant, or alternative recall pathways may remain undetected if they are not activated under the evaluated conditions (Canby et al., 2025). BCTI therefore establishes disruption of the tested causal support for factual recall rather than complete parameter-level deletion.

Accordingly, causal restoration reduction is best viewed as an intervention-specific measure of mechanistic disruption. It asks whether states that restored the target answer before unlearning can still do so under the same patching protocol. Comparing this measure with behavioral forgetting can distinguish two failure modes: behavioral suppression despite preserved causal support, and weakened measured support despite residual access through untested pathways. This joint interpretation motivates reporting both quantities rather than treating either measure as standalone evidence of complete erasure.

The current experiments are further limited to independently unlearned factual associations in two 7B-scale model families, Pythia-6.9B-deduped (Biderman et al., 2023) and OLMo-7B. Related or jointly unlearned facts may share causal components, increasing conflicts between forgetting and retention gradients (Wang et al., 2025; Yu et al., 2020). Moreover, localization and verification require repeated corrupted and activation-patched forward passes, so parameter-efficient LoRA updates do not eliminate the tracing overhead (Hu et al., 2022). Future work should develop more efficient causal estimation, examine stability after relearning or continued training, and extend evaluation to larger models, multilingual and compositional knowledge, dependent fact sets, and large-scale deletion requests (Dorna et al., 2025).

7 Conclusion

We propose BCTI, a bidirectional causal tracing and intervention framework for factual knowledge unlearning. Unlike methods that define forgetting solely through output changes, BCTI treats the causal contributions supporting target knowledge recall as the primary object of intervention. It uses forward and reverse causal tracing to localize internal components that jointly support factual retrieval and directly weakens their ability to restore the target answer.

The results in Table 1 show that BCTI achieves substantial answer-level forgetting while limiting damage to non-target knowledge, and the causal audit in Table 2 confirms that this behavioral change accompanies a large causal restoration effect reduction. The same ablations further show that reliable forgetting benefits from both accurately identifying the relevant recall pathways and directly weakening their causal contributions.

Overall, our results support the claim that weakening the causal support underlying target knowledge recall supports a more reliable assessment of forgetting than behavioral suppression alone.

AI USE STATEMENT

In this work, we used generative AI tools primarily for language editing and polishing, including improving grammar, clarity, readability, academic phrasing, and the presentation of author-written text. We also used generative AI tools for technical assistance, such as helping diagnose implementation errors and suggesting possible fixes during code development.

We have not used generative AI tools to independently develop the core research ideas, design the proposed method, determine the experimental setup, generate experimental results, or make the scientific conclusions of this work. These aspects were primarily developed and carried out by the authors.

Additionally, we used generative AI tools to assist with literature search, information retrieval, and the organization and synthesis of relevant background materials. AI-generated suggestions and summaries were used only as auxiliary references. Relevant technical claims, cited works, and literature-based statements were checked by the authors against the corresponding sources before being included in the manuscript.

We have reviewed all AI-assisted work. The authors made the final decisions regarding the research ideas, methodology, experimental design, analysis, interpretation, and presentation of the results, and take responsibility for the final content of this work, including text, claims, code, figures, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, the collection of private user data, or personally identifiable information. Our experiments are conducted on publicly available language models and controlled factual associations. Therefore, no institutional review board approval was required for the experiments reported in this paper.

Machine unlearning can contribute to privacy protection, removal of outdated or incorrect information, and compliance with data-deletion requirements. However, the same techniques may also be misused to selectively suppress legitimate information, conceal undesirable model behavior, or create misleading claims that particular knowledge has been completely removed.

To avoid overstating the guarantees provided by our method, we interpret the reduction in causal restoration scores only as evidence that the evaluated recall pathways have been weakened, rather than as proof of complete parameter-level deletion of the target knowledge.

The proposed causal localization and intervention framework may also affect model behavior beyond the intended target if the selected internal components support multiple facts or capabilities. We therefore explicitly evaluate retention performance and discuss this limitation in the paper. More broadly, practical deployment of machine unlearning systems should consider privacy, safety, fairness, legal compliance, and the possibility of unintended or malicious use.

REPRODUCIBILITY STATEMENT

We provide the source code and reproduction materials at `https://anonymous.4open.science/r/BCTI-MU-863E`. Section 5.1 specifies the evaluated models, data, baselines, optimization setup, and reported metrics. Appendix A.1 documents the prompt-disjoint fit, validation, and test construction; Appendix A.2 gives the localization, calibration, optimization, checkpoint-selection, and evaluation rules; and Appendices B and C report the detailed target-level and cross-model results. Together, these resources provide the implementation and experimental specifications needed to reproduce the reported results.

REFERENCES

Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Stickland, Tomek Korbak, and Owain Evans. The reversal curse: Llms trained on "a is b"fail to learn "b is a". In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 18623–18642,

2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/5178b2f2d7c44aa390c0777dc77b3f0c-Paper-Conference.pdf.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 23–29 Jul 2023.

Marc E. Canby, Adam Davies, Chirag Rastogi, and Julia Hockenmaier. How reliable are causal probing interventions? In Kentaro Inui, Sakriani Sakti, Haofen Wang, Derek F. Wong, Pushpak Bhattacharyya, Biplab Banerjee, Asif Ekbal, Tanmoy Chakraborty, and Dhirendra Pratap Singh (eds.), *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 857–878, Mumbai, India, December 2025. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. ISBN 979-8-89176-298-5. URL https://aclanthology.org/2025.ijcnlp-long.47/.

Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, Zico Kolter, Zachary Lipton, and Pratyush Maini. Openunlearning: Accelerating llm unlearning via unified benchmarking of methods and metrics. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-1458. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/3e4a38f228427ab819ba7899003a44b1-Paper-Datasets_and_Benchmarks_Track.pdf.

Jiacheng Du, Zhibo Wang, Jie Zhang, Xiaoyi Pang, Jiahui Hu, and Kui Ren. Textual unlearning gives a false sense of unlearning. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 14579–14597. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/du25d.html.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for llm unlearning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 1540–1567. Curran Associates, Inc., 2025. doi: 10.52202/085713-0058. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/02443e4e008231e0af0855f3cc70ed17-Paper-Conference.pdf.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235. Association for Computational Linguistics, December 2023. URL https://aclanthology.org/2023.emnlp-main.751/.

Phillip Huang Guo, Aaquib Syed, Abhay Sheshadri, Aidan Ewart, and Gintare Karolina Dziugaite. Mechanistic unlearning: Robust knowledge unlearning and editing via mechanistic localization. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 20964–20992. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/guo25k.html.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko,

M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 17643–17668. Curran Associates, Inc., 2023. doi: 10.52202/075280-0774. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/3927bbdcf0e8d1fa8aa23c26f358a281-Paper-Conference.pdf.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Shengyuan Hu, Yiwei Fu, Steven Wu, and Virginia Smith. Unlearning or obfuscating? jogging the memory of unlearned llms via benign relearning. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 8857–8888, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/18fd48d9cbbf9a20e434c9d3db6973c5-Paper-Conference.pdf.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14389–14408. Association for Computational Linguistics, July 2023. URL https://aclanthology.org/2023.acl-long.805/.

Zhuoran jin, Pengfei Cao, Chenhao Wang, Zhitao He, Hongbang Yuan, Jiachun Li, Yubo Chen, Kang Liu, and Jun Zhao. Rwku: Benchmarking real-world knowledge unlearning for large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 98213–98263. Curran Associates, Inc., 2024. doi: 10.52202/079017-3117. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/b1f78dfc9ca0156498241012aec4efa0-Paper-Datasets_and_Benchmarks_Track.pdf.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, July 2020. URL https://aclanthology.org/2020.acl-main.428/.

Bo Liu, Qiang Liu, and Peter Stone. Continual learning and private unlearning. In Sarath Chandar, Razvan Pascanu, and Doina Precup (eds.), *Proceedings of The 1st Conference on Lifelong Learning Agents*, volume 199 of *Proceedings of Machine Learning Research*, pp. 243–254. PMLR, 22–24 Aug 2022. URL https://proceedings.mlr.press/v199/liu22a.html.

Yujian Liu, Yang Zhang, Tommi Jaakkola, and Shiyu Chang. Revisiting who's harry potter: Towards targeted unlearning from a causal intervention perspective. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8708–8731. Association for Computational Linguistics, November 2024. doi: 10.18653/v1/2024.emnlp-main.495. URL https://aclanthology.org/2024.emnlp-main.495/.

Tyler Lizzo and Larry Heck. UNLEARN efficient removal of knowledge in large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 7272–7283. Association for Computational Linguistics, April 2025. ISBN 979-8-89176-195-7. URL https://aclanthology.org/2025.findings-naacl.405/.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Guangzhi Sun, Potsawee Manakul, Xiao Zhan, and Mark Gales. Unlearning vs. obfuscation: Are we truly removing knowledge? In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 11457–11467. Association for Computational Linguistics, November 2025. ISBN 979-8-89176-332-6. URL https://aclanthology.org/2025.emnlp-main.577/.

Bozhong Tian, Xiaozhuan Liang, Siyuan Cheng, Qingbin Liu, Mengru Wang, Dianbo Sui, Xi Chen, Huajun Chen, and Ningyu Zhang. To forget or not? towards practical knowledge unlearning for large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1524–1537. Association for Computational Linguistics, November 2024. URL https://aclanthology.org/2024.findings-emnlp.82/.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Qicheng Wang, Quanhe Wang, and Lu Wang. Exact: Expert attribution via causal tracing for targeted knowledge unlearning in mixture-of-experts language models. In De-Shuang Huang, Chuanlei Zhang, Wei Chen, and Bo Li (eds.), *Advanced Intelligent Computing Technology and Applications*, pp. 170–181, Singapore, 2027. Springer Nature Singapore.

Qizhou Wang, Jin Zhou, (Andrew) Zhanke Zhou, Saebyeol Shin, Bo Han, and Kilian Weinberger. Rethinking llm unlearning objectives: A gradient perspective and go beyond. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 61897–61931, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/9ad307ab459fb93163248a0a56959307-Paper-Conference.pdf.

Hao-Xiang Xu, Jun-Yu Ma, Zhen-Hua Ling, Quan Liu, Cong Liu, and Jia-Chen Gu. Evaluating the reversal curse in model editing. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=jAHwodCUxP.

Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. Machine unlearning of pre-trained large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8403–8419, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.457/.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 4(2):100211, 2024b. ISSN 2667-2952. doi: https://doi.org/10.1016/j.hcc.2024.100211. URL https://www.sciencedirect.com/science/article/pii/S266729522400014X.

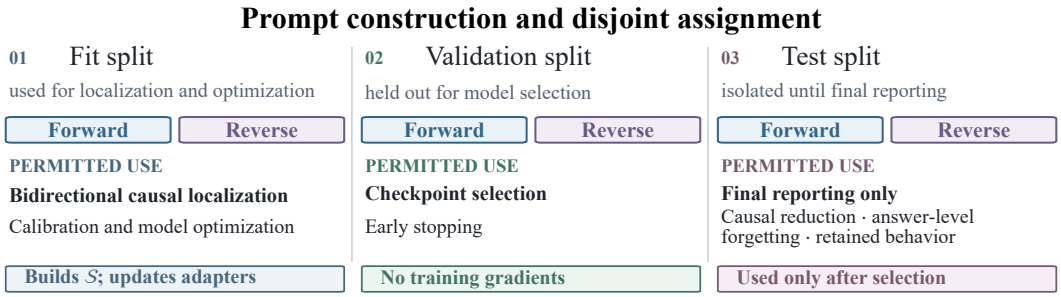


Figure 6: Prompt construction and disjoint fit, validation, and test assignment in BCTI. The split is fixed before localization so final audits use prompts unseen during intervention construction and checkpoint selection.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 5824–5836. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/3fe78a8acf5fda99de95303940a2420c-Paper.pdf.

Naixin Zhai, Pengyang Shao, Binbin Zheng, Yonghui Yang, Fei Shen, Long Bai, and Xun Yang. Maximizing local entropy where it matters: Prefix-aware localized LLM unlearning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19514–19531. Association for Computational Linguistics, July 2026. ISBN 979-8-89176-390-6. URL https://aclanthology.org/2026.acl-long.893/.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 1651–1678, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/06a52a54c8ee03cd86771136bc91eb1f-Paper-Conference.pdf.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=MXLBXjQkmb.

A Additional Experimental Details and Analysis

This appendix first clarifies the prompt-disjoint protocol and the fixed implementation rules underlying BCTI, then reports target-level behavioral, causal, and retention results on Pythia-6.9B, and finally reports results on OLMo-7B. All analyses use the definitions and data summarized in the main paper.

A.1 Prompt-disjoint construction and evaluation

Prompts are assigned to disjoint fit, validation, and test splits before localization (Figure 6). Fit prompts are the only prompts used to construct the intervention set, calibrate directional weights, and optimize adapters.

Validation first applies a hard gate using the larger remaining directional restoration score, so improvement in one direction cannot hide failure in the other; within the causal-feasible set, checkpoint selection then prioritizes feasibility, retention safety, weak-direction forgetting, and finally the causal score. Test prompts remain isolated until final behavioral, causal, and retention evaluation. This ordering prevents the causal audit from reusing prompts that determined either the intervention sites or the selected checkpoint.

A.2 Localization, calibration, and gradient details

The following fixed rules instantiate the operations summarized in Section 4.2. They use fit prompts only; neither validation nor test prompts enter localization, directional calibration, or gradient construction.

Robust localization. At each residual-stream input, attention output, and MLP output, normalized restoration values are clipped to $[0, 1]$. For each direction d, let μ_d and m_d be their mean and median, and let κ_d be the fraction reaching the causal-effect threshold 0.20. We aggregate them as $R_d = \frac{1}{2}(\mu_d + m_d)\frac{1}{2}(1 + \kappa_d)$, average site scores within each layer, and apply the harmonic mean to the resulting forward and reverse scores. The candidate set is the union of the top 12 bidirectional layers and the top 6 layers in each direction. Candidates must reach 50% of the best bidirectional score unless this leaves fewer than four layers. We select four layers, reserving at least one per direction with causal coverage of at least 0.33, and fill the remaining slots by bidirectional score; retain restoration is used only to prefer the safer layer when eligible candidates are otherwise tied.

Directional calibration. Eligible fit prompts are partitioned by direction into optimization and calibration subsets, reserving one calibration prompt per direction whenever possible (a 25% fraction capped at one). Every four updates, the relative causal reduction r_d on these untouched fit prompts defines directional difficulty $a_d = \max(0.90 - r_d, \varepsilon)$. The two difficulties are normalized to mean one and clipped to $[0.5, 1.5]$ to obtain λ_d, so the direction with less causal progress receives more weight. When a direction has fewer than two optimization prompts, this estimate is blended with its initial fit-trace difficulty; validation and test measurements are never used.

Gradient composition. The causal, answer, and retention coefficients are all 1.0, but the update is hierarchical. If the answer gradient conflicts with the causal gradient, its conflicting component is projected out; its remaining norm is capped at 0.90 of the causal-gradient norm and is disabled when the causal gradient is zero. The resulting forgetting gradient is then combined with the retention gradient using module-level PCGrad (Yu et al., 2020): for a module whose retention-gradient norm is at least 1% of its forgetting-gradient norm, a negative inner product triggers projection of the forgetting gradient away from the retention gradient before the latter is added. The final combined gradient is clipped to norm 0.5.

Evaluation eligibility. On Pythia-6.9B, a forget prompt contributes to directional behavioral metrics only when the frozen base model greedily generates the complete target answer. Applying this rule leaves at least one eligible forward test prompt for 47 target runs and at least one eligible reverse prompt for 46; the remaining three and four runs, respectively, are omitted only from the corresponding directional summaries. A causal trace is valid only when its clean–corrupted probability gap exceeds 10^{-8}. For OLMo-7B, whose base model can assign substantial probability without reproducing the complete surface form, eligibility additionally accepts a base answer score of at least 0.01; the same two-direction and held-out evaluation structure is otherwise retained.

B Detailed Pythia-6.9B Results

Table 3 summarizes held-out test results across the 50 target-specific runs. Figure 7 complements the aggregate statistics with the target-level forgetting and retention matrix used to expose heterogeneous targets and retention outliers.

B.1 Behavioral forgetting and directional asymmetry

Under the eligibility rule above, forward forgetting is $91.69\% \pm 8.40\%$ over 47 targets, whereas reverse forgetting is $85.20\% \pm 10.56\%$ over 46 targets. This gap explains why aggregate Forget is accompanied by Min-dir.: an average can otherwise conceal a weaker inverse-query route.

Figure 8 resolves this asymmetry by target and relation category. Because the 50 targets contain unequal numbers of capital, location, and currency relations, the category panels are descriptive rather than controlled relation-level comparisons. Their purpose is to expose dispersion and directional failure modes.

Table 3: Test-set results of BCTI on Pythia-6.9B.

Test metric	Mean ± SD	Median	Range
Answer-score reduction	87.77% ± 7.77%	89.45%	64.04%–99.81%
Minimum-direction reduction	83.28% ± 10.37%	86.10%	54.49%–99.81%
Causal restoration reduction	98.58% ± 1.51%	98.95%	92.04%–99.95%
Minimum-direction causal reduction	97.72% ± 2.38%	98.63%	86.77%–99.91%
Retain KL	0.00649 ± 0.00426	0.00511	0.00127–0.02103
Maximum retain answer drop	12.47% ± 10.74%	9.40%	4.11%–57.47%

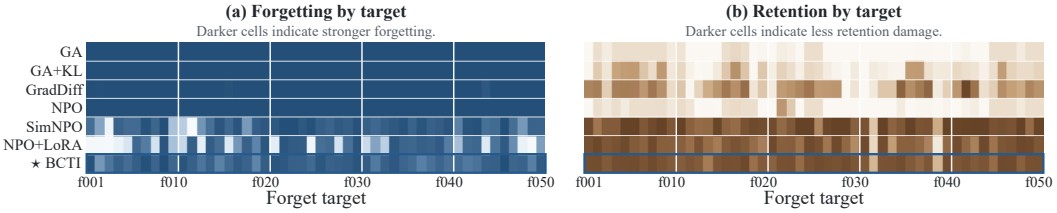


Figure 7: Target-level BCTI results on Pythia-6.9B. (a) Answer-score reduction across forget targets; darker cells denote stronger forgetting. (b) Retention behavior; darker cells denote less retention damage.

B.2 Causal disruption and retention heterogeneity

Figure 9 places behavioral forgetting beside causal reduction and retention damage. Behavioral reduction spans 64.04%–99.81%, whereas causal-restoration reduction remains within 92.04%–99.95%. Their different ranges show that consistently large reductions in audited restoration do not imply uniform output suppression. Maximum retain-answer damage also has a pronounced upper tail, showing why retain KL alone is insufficient.

Across the 50 runs, mean causal-restoration reduction is $98.58\% \pm 1.51\%$, and minimum-direction causal reduction is $97.72\% \pm 2.38\%$. The remaining maximum directional causal score is 0.0020 ± 0.0023. Figure 10 shows attenuation across targets and selected layers rather than an effect driven by one target or a single isolated layer. This is evidence for disruption of the audited pathways, not proof that every representation or alternative route has been erased.

Mean retain KL is 0.00649 ± 0.00426, but mean maximum retain-answer drop is $12.47\% \pm 10.74\%$ and reaches 57.47% for the largest outlier. Similar forgetting can therefore coexist with different localized retention costs. The paired metrics distinguish average distributional stability from the worst observed factual degradation.

C Evaluation on OLMo-7B

We repeat BCTI on OLMo-7B to test whether the Pythia result is specific to one model parameterization. Table 4 reports the same held-out metrics. Mean causal-restoration reduction is $97.81\% \pm 1.94\%$, and the remaining maximum directional causal score is 0.00184 ± 0.00208, reproducing strong disruption of the audited pathway in a second model family.

Behavioral forgetting is lower than on Pythia-6.9B: mean answer-score reduction differs by 5.36 percentage points and Min-dir. by 5.17 points. On OLMo-7B, forward forgetting is $87.75\% \pm 12.51\%$, compared with $81.82\% \pm 10.81\%$ in reverse, so directional asymmetry persists across the two families.

Retention damage is lower on average, with retain KL 0.00275 ± 0.00160 and mean maximum retain-answer drop $10.08\% \pm 8.37\%$, although the worst-case tail remains non-negligible.

Checkpoint selection differs across models: the selected checkpoint occurs at step 66.64 ± 9.62 for Pythia-6.9B and 54.40 ± 15.96 for OLMo-7B. Figure 11 shows causal reduction approaching sat-

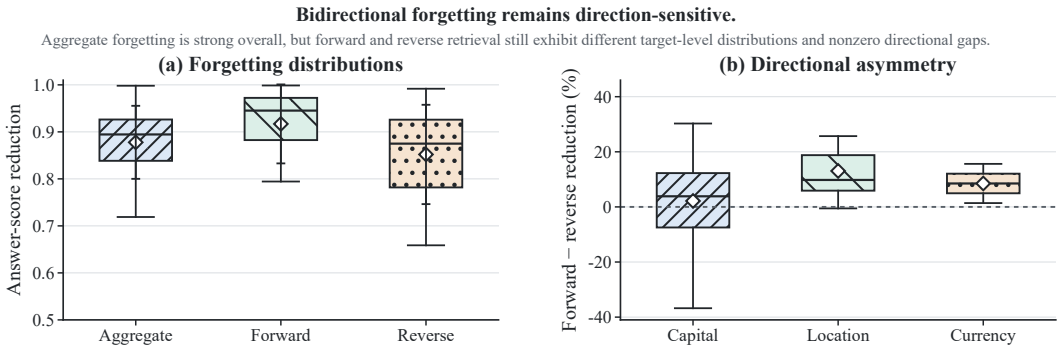


Figure 8: Bidirectional forgetting on Pythia-6.9B. (a) Distributions of aggregate, forward, and reverse answer-score reduction. (b) Forward-minus-reverse reduction across factual-relation categories; the dashed line denotes equal reduction.

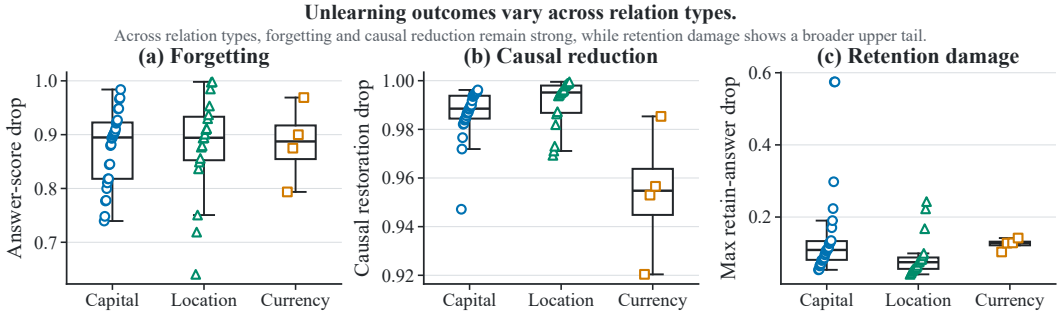


Figure 9: Descriptive distributions across factual-relation categories on Pythia-6.9B. (a) Answer-score reduction. (b) Causal-restoration reduction. (c) Maximum retain-answer drop.

uration before all behavioral and retention quantities stabilize. This supports selecting checkpoints on held-out causal criteria rather than automatically using the final step.

C.1 Cross-model observations

Agreement across models is strongest for causal-restoration reduction, whereas behavioral forgetting and retention remain more target- and model-dependent. This replication shows that the measured mechanistic effect is also observed in both evaluated 7B-scale families, but it does not establish architecture- or scale-independent guarantees. Behavioral forgetting, mechanistic disruption, and retention should therefore be reported as complementary dimensions rather than collapsed into one success criterion.

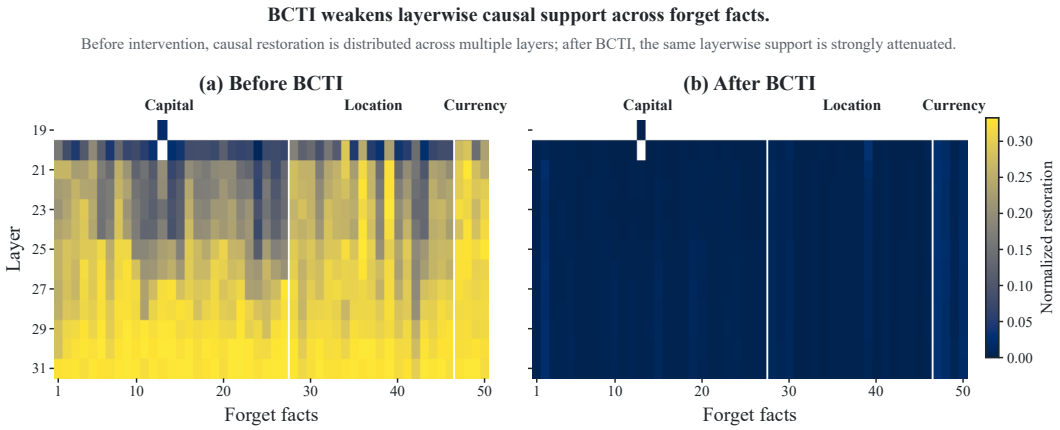


Figure 10: Target-by-layer normalized causal restoration on Pythia-6.9B before and after BCTI. Rows denote transformer layers, columns denote target facts, and white cells indicate unavailable values.

Table 4: Test-set results of BCTI on OLMo-7B.

Test metric	Mean ± SD	Median	Range
Answer-score reduction	82.42% ± 9.54%	83.59%	55.72%–97.10%
Minimum-direction reduction	78.11% ± 11.47%	80.55%	36.03%–96.01%
Causal restoration reduction	97.81% ± 1.94%	98.48%	91.40%–99.87%
Minimum-direction causal reduction	96.42% ± 3.33%	97.60%	83.48%–99.86%
Retain KL	0.00275 ± 0.00160	0.00225	0.00085–0.00752
Maximum retain answer drop	10.08% ± 8.37%	7.92%	2.22%–48.04%

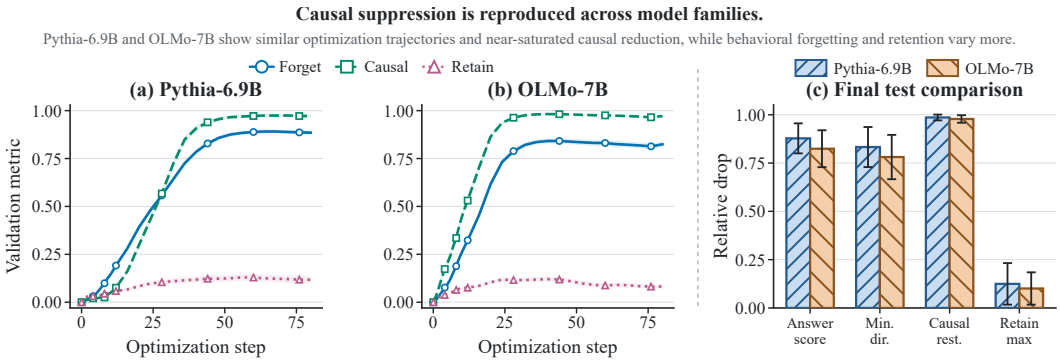


Figure 11: Optimization dynamics and cross-model comparison. (a,b) Validation trajectories on Pythia-6.9B and OLMo-7B. (c) Final test comparison of answer-score reduction, Min-dir. reduction, causal-restoration reduction, and maximum retain-answer drop.