WHEN UNLEARNED MODELS REASON, WHERE DOES THE ANSWER COME FROM?

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine unlearning is evaluated by asking an unlearned model about the removed information, but if the model is allowed to reason first, it often answers these forgotten questions correctly again. However, existing evaluations of unlearning only compare accuracy with and without reasoning or inspect reasoning traces for leaked answers, so they cannot show whether the final answer takes its answer from the trace. We instead treat the reasoning trace as an object of intervention, storing and editing it before a separate answering step. Unlike accuracy comparisons and leakage checks, these edits show where the final answer takes its answer from. Concretely, following studies of factual recall, we compare reasoning with a dummy trace, a meaningless text of the same length, and split the recovery measured against it into a dummy penalty and a gain over direct answering. We also mask the answer span against equal-size placebo masks, substitute another author's name of equal token length, and give identical traces to different checkpoints. Experiments on TOFU with seven unlearning methods show that most of the recovery measured against the dummy trace comes from the dummy trace lowering accuracy, not from reasoning. In smaller models, the final answer takes whatever answer the trace states, even in a retrained model that never saw the facts, and unlearning changes what the trace says rather than how the answer uses it. In larger models, unlearning also weakens how much the final answer uses the trace.

1 INTRODUCTION

Machine unlearning aims to remove specific information from a trained model so that it behaves as if it had never seen that information. It is proposed for private data (Maini et al., 2024), copyrighted text (Shi et al., 2025), and hazardous knowledge (Li et al., 2024), and successful unlearning is commonly judged by whether the unlearned model can still answer questions about what was removed. However, several studies (Yoon et al., 2025; Wang et al., 2025; Le et al., 2026) find that once an unlearned model is allowed to reason before answering, the forgotten information comes back in its reasoning. This observation raises the following question.

Research Question

When an unlearned model answers a forgotten question correctly after reasoning, where does the answer come from? Is the recovery only an artifact of the comparison, does reasoning help the model reach knowledge it cannot state directly, or does the model write the answer into its own reasoning trace?

Answering this question is hard for three reasons. First, the measured recovery depends on the comparison, and a control such as a meaningless text of the same length can itself change accuracy. Second, an answer found in a reasoning trace need not be the source of the final answer, which the model might give without the trace. Third, each checkpoint writes its own traces, so checkpoints can differ either in what they write or in how they use it. Inspired by chain-of-thought faithfulness tests that edit the reasoning and check whether the answer changes (Lanham et al., 2023), we store each reasoning trace and edit it before a separate answering step, as illustrated in Figure 2. We compare reasoning with direct answering as well as with the dummy trace used in studies of factual

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.

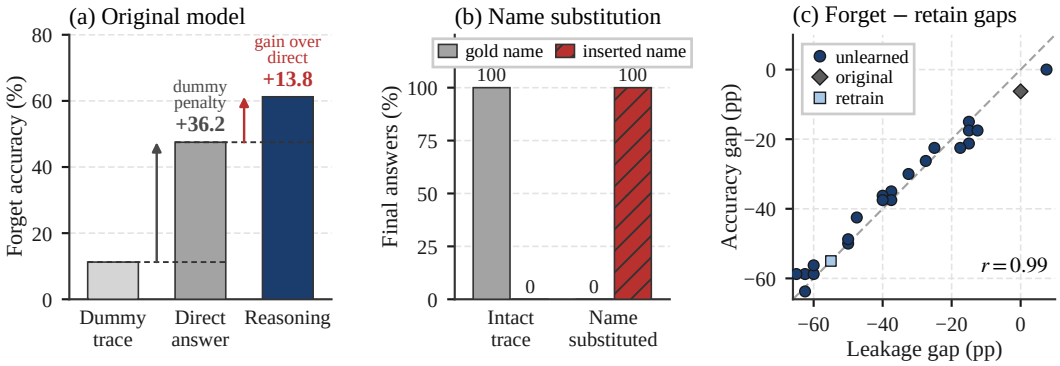


Figure 1: **Where the recovered answer comes from.** Most of the apparent recovery is a baseline artifact, and the rest is the answer that the reasoning trace supplies. **(a)** The dummy trace lowers accuracy far below direct answering, which accounts for most of the apparent recovery. **(b)** Replacing the name in the trace makes every final answer name the inserted person. **(c)** Across checkpoints, the forget−retain gap in accuracy with reasoning matches the gap in answer leakage.

recall (Gekhman et al., 2026); we mask the answer against equal-size placebo masks and, as in knowledge-conflict research (Longpre et al., 2021), substitute another person's name; and we give identical traces to different checkpoints. Each of these designs removes one of the three confounds. Unlike existing evaluations, which score the reasoning trace and the final answer separately, they test the link between the two, and this link shows where the recovered answer comes from.

We study this question on TOFU (Maini et al., 2024), using the original model, a retrained reference, and unlearned checkpoints from seven methods (Dorna et al., 2025), and we repeat the tests on models of different parameter sizes. As shown in Figure 1, we first study how much of the apparent recovery is real, and find that most of it is an artifact of the comparison, much as apparent emergent abilities can be an artifact of the metric (Schaeffer et al., 2023). We further study where the real recovery comes from, and find that it comes from the reasoning trace, since the final answer takes whatever answer the trace states, even in a retrained model that never saw the facts. Finally, we study what unlearning changes along this path, and find that in smaller models it changes only how often the trace states the answer, whereas in larger models it also weakens how much the final answer uses the trace. For unlearning, this means that recovery should be measured against direct answering, and that a method has to stop the model from writing the answer, not only from stating it when asked directly. We make three contributions.

- We decompose reasoning-time recovery by splitting it into a dummy penalty and a gain over direct answering, and by separating answer leakage from answer use with interventions on stored reasoning traces.
- We find that most of the apparent recovery is an artifact of the comparison, and that the real recovery is the answer that the model writes into its reasoning trace and then copies.
- We find that unlearning shows up in what the trace says. In smaller models, it changes how often the trace states the answer, not how the final answer uses the trace, and in larger models it also weakens that use.

2 Related Work

LLM unlearning. Unlearning methods fine-tune a model against a forget set or perturb its representations (Maini et al., 2024; Zhang et al., 2024; Li et al., 2024), and benchmarks score them on fictitious biographies, hazardous knowledge, and copyrighted text (Maini et al., 2024; Li et al., 2024; Shi et al., 2025; Dorna et al., 2025). Removed answers nevertheless resurface after quantization, light fine-tuning, activation edits, and jailbreaks (Zhang et al., 2025; Hu et al., 2025; Deeb & Roger, 2024; Łucki et al., 2024; Lynch et al., 2024). These studies infer that the knowledge is still stored in the weights, and careful designs ensure that the attack itself cannot supply the an-

swer (Deeb & Roger, 2024). Reasoning removes this safeguard, because the model writes its own context. Whether that context helps also depends on a second ability, using supplied knowledge, which unlearning can degrade (Peng et al., 2026). An answer recovered after reasoning therefore needs a test before it counts as stored knowledge.

Reasoning in unlearning. Work on unlearning reasoning models treats the chain of thought as a new place where forgotten content leaks. Answer-level objectives leave that content in the trace, and decoding with little thinking can still reveal it (Yoon et al., 2025; Wang et al., 2025). Methods for these models accordingly unlearn the trace or build reasoning into the unlearning target (Li et al., 2026; Le et al., 2026; Liao et al., 2026). Mechanistic studies instead suggest that reasoning helps a model recall facts, through a computational buffer and factual priming (Gekhman et al., 2026) or by re-deriving removed facts from retained ones (Wu et al., 2025). Neither line tests whether the final answer of an unlearned model takes its answer from the reasoning trace or produces it without the trace. We test this with interventions from chain-of-thought faithfulness (Lanham et al., 2023; Pfau et al., 2024; Tutek et al., 2025; Mitsuhashi et al., 2026) and knowledge-conflict research (Longpre et al., 2021; Xie et al., 2024). This test gives trace-level unlearning a concrete target. In Llama-3.2-1B, recovery tracks whether the answer is written, because the final answer uses what is written.

3 Attributing reasoning-time recovery

This section turns the attribution question into quantities we can measure. We study reasoning in a generate-then-read interface (Yu et al., 2023; Sun et al., 2023), in which the model first writes a short reasoning trace and a separate step answers the question with that trace as context. Recovery in this interface can come from four sources. Two lie outside the answer itself, namely the comparison baseline and help that the trace gives without stating the answer, such as a computational buffer or factual priming (Gekhman et al., 2026). The other two concern the answer itself. The trace may state the answer and the answering step take it from there, or the answering step may produce the answer as it would without reasoning. The two differ in whether the final answer draws on contextual or parametric knowledge (Longpre et al., 2021; Xie et al., 2024). We separate these sources in three steps. First, we compare reasoning with direct answering as well as with a dummy trace. Second, we record how often traces state the answer. Third, we edit stored traces to test whether the final answer uses what they state, first within each checkpoint and then on identical traces across checkpoints.

Notation. A checkpoint s answers factual questions from a forget set F and a retain set T. We compare three contexts, namely none (*direct* answering, D), a *dummy* trace of the same length that says nothing about the question (M), and the checkpoint's own reasoning trace (N). $A_c(s, z)$ denotes the accuracy of the final answer under context c on set $z \in \{F, T\}$.

Baseline decomposition. A natural control for reasoning-time recovery is the dummy trace. It matches the length of a reasoning trace but says nothing about the question. Comparing against it is meant to separate what reasoning says from the computational buffer it provides (Gekhman et al., 2026; Pfau et al., 2024). We add direct answering as a second baseline. The two baselines split the apparent recovery exactly into two terms,

$$\underbrace{A_{\mathrm{N}} - A_{\mathrm{M}}}_{\text{apparent recovery}} = \underbrace{A_{\mathrm{N}} - A_{\mathrm{D}}}_{\text{gain over direct}} + \underbrace{A_{\mathrm{D}} - A_{\mathrm{M}}}_{\text{dummy penalty}}. \tag{1}$$

The first term is the gain that reasoning brings over answering without it, and the second is the penalty that the dummy trace itself imposes. The dummy comparison therefore measures the content of reasoning only if this penalty is close to zero; Section 4.2 shows that it is not.

Leakage and use. *Answer leakage* $E(s, z)$ is the share of reasoning traces that state the answer; *answer use* is the extent to which the final answer follows an answer that the trace states. Trace-level evaluations of unlearning measure such leakage (Yoon et al., 2025; Wang et al., 2025); our released tables call it exposure. Let $U^+(s, z)$ and $U^-(s, z)$ be the final accuracy on traces that do and do not state it. Then

$$A_{\mathrm{N}}(s, z) = E(s, z)\, U^+(s, z) + \big(1 - E(s, z)\big)\, U^-(s, z). \tag{2}$$

Question (Forget Set). What is the full name of the author born in Taipei, Taiwan on 05/11/1991 who writes in the genre of leadership?

Original Model's Reasoning Trace. The author's full name is Hsiao Yun-Hwa, and she was born in Taipei, Taiwan on 05/11/[P1]1991. She specializes in the genre of leadership and has written extensively on[P2] the subject.

Edit	Six Tokens Replaced	Replaced By	Final Answer If It Comes From	
			The Trace	Direct Answering
Controls (Both Sources Predict the Same Final Answer)				
None (Intact)	–	–	Correct	Correct
Placebo Mask P1	on 05/11/	X X X X X X	Correct	Correct
Placebo Mask P2	leadership … on	X X X X X X	Correct	Correct
Tests (The Two Sources Predict Different Final Answers)				
Answer Mask	Hsiao Yun-Hwa	X X X X X X	Wrong	Correct
Entity Substitution	Hsiao Yun-Hwa	Aysha Al-Hashim	Aysha Al-Hashim	Hsiao Yun-Hwa

Figure 2: **Four edits to a reasoning trace.** Only the answer mask and entity substitution change the answer that the trace states, so only they tell whether the final answer comes from the trace; the placebo masks control for damage to the trace.

Equation (2) is an accounting identity, yet the two ways of producing the answer predict different values of its terms. If the final answer takes the answer from the trace, answer use is high, so $U^+ \gg U^-$ and editing the answer in a trace changes the final answer. Forgetting under reasoning is then governed by answer leakage, that is, by what the checkpoint writes. If the model produces the answer as it would without reasoning, answer use is low, so $U^+ \approx U^-$ and editing the trace matters little.

Masking and substitution. As in chain-of-thought faithfulness tests (Lanham et al., 2023), we test answer use directly by storing traces that state the answer and editing them before the final answer. *Answer masking* replaces the tokens of the answer; two *placebo* masks, P1 and P2, replace the same number of tokens elsewhere in the trace and leave the answer intact, as illustrated in Figure 2. On a set $\mathcal{S}$ of such traces, the *masking effect* is

$$C(s, z; \mathcal{S}) = \tfrac{1}{2}\big[A_{\mathrm{P1}}(s, z; \mathcal{S}) + A_{\mathrm{P2}}(s, z; \mathcal{S})\big] - A_{\mathrm{mask}}(s, z; \mathcal{S}). \tag{3}$$

Because the placebos damage the trace as much as the answer mask does, C separates answer use from damage to the trace; our released tables call it the carrier effect. A second intervention asks which answer is used. Following knowledge-conflict research (Longpre et al., 2021), *entity substitution* replaces every mention of the answer with an incompatible name of equal token length. The final answer then names the gold person, the inserted person, or neither.

Masking interaction. To compare answer use across checkpoints without confounding it with trace generation, all checkpoints answer from identical traces, one set $\mathcal{S}_O$ written by a single checkpoint. We then estimate the reference- and split-calibrated *masking interaction*

$$\Gamma(s, r) = \big[C(s, F; \mathcal{S}_O) - C(r, F; \mathcal{S}_O)\big] - \big[C(s, T; \mathcal{S}_O) - C(r, T; \mathcal{S}_O)\big], \tag{4}$$

where r is a reference checkpoint. $\Gamma \approx 0$ means that unlearning left answer use unchanged, beyond any change shared with the retain set.

4 Experiments

In this section, we answer the following questions through experiments. **Q1.** How much of the apparent recovery is a baseline artifact? **Q2.** Does the gain over direct answering come from the reasoning trace? **Q3.** Does unlearning change what the trace says, or whether the final answer follows it? **Q4.** Do these answers hold for models of different parameter sizes? Section 4.1 first describes the experimental setup, and Figure 3 shows the design. More details, including the checkpoints, the construction of each context, decoding, and the interventions, are in Appendix A.

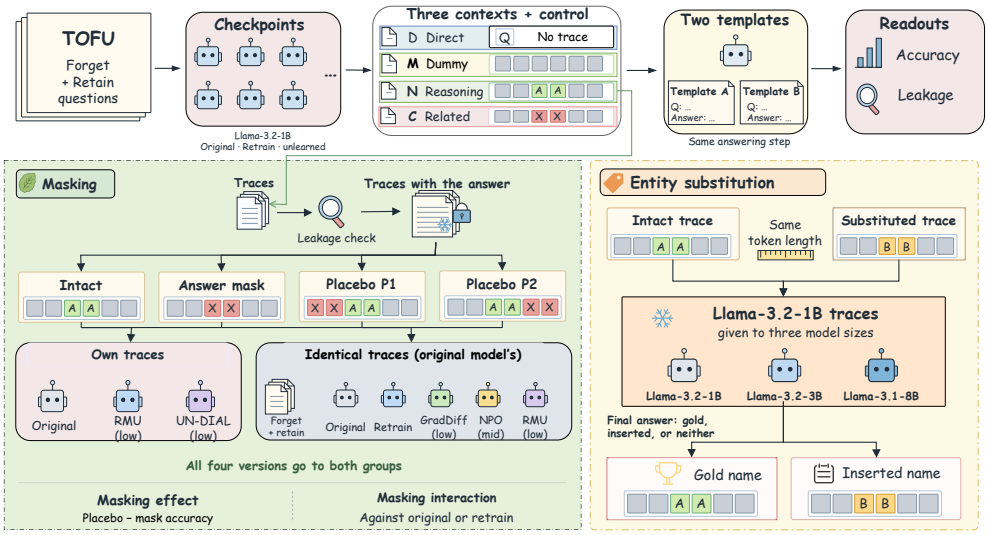


Figure 3: **Experimental design.** Each comparison changes one factor and holds the rest fixed, so a change in the final answer can be traced to that factor. In the top row, the contexts separate the dummy penalty from the gain over direct answering. In the bottom row, masking tests whether the final answer uses the answer in the trace, and entity substitution tests which answer it uses.

Key Findings

- **Where does the apparent recovery come from?** Mostly from the comparison. The dummy trace lowers accuracy below answering directly, so a comparison against it overstates what reasoning adds (Section 4.2).
- **What does reasoning add over answering directly?** The answer itself. The final answer takes whatever answer the trace states, so a correct answer after reasoning shows what the trace contains, not what the model would answer directly (Section 4.3).
- **What does unlearning change?** Mainly what the trace says. Unlearning lowers how often the trace states the answer; as parameter size grows, it also weakens how much the final answer uses the trace (Sections 4.4 and 4.5).

4.1 Experimental setup

Dataset. We use TOFU (Maini et al., 2024), whose questions concern fictitious authors, so a model can know the answers only from fine-tuning on TOFU or from its context. The evaluation set has 40 forget-split (`forget10`) and 40 retain-split questions, two atomic-answer questions for each of 20 authors per split.

Models. We run most experiments on Llama-3.2-1B (Grattafiori et al., 2024), and we use Llama-3.2-3B and Llama-3.1-8B to test whether the results change with parameter size. We also test a model from a different family, Qwen2.5-1.5B (Yang et al., 2024). All four are instruction-tuned versions. Each model has two reference checkpoints. The original model is fine-tuned on all TOFU authors, and the retrain model is fine-tuned on the retain authors only (`retain90`). The retrain model never sees the forget facts and marks what ideal unlearning would achieve.

Unlearning methods. We study GradDiff and IdkDPO (Maini et al., 2024), NPO (Zhang et al., 2024), SimNPO (Fan et al., 2025), RMU (Li et al., 2024), UN-DIAL (Dong et al., 2025), and AltPO (Mekala et al., 2025), seven unlearning methods implemented in OpenUnlearning (Dorna et al., 2025). On Llama-3.2-1B, each method is applied at low, mid, and high unlearning depth (Table 1; Appendix A). The other models are unlearned with RMU, GradDiff, and NPO.

Table 1: **Where the apparent recovery comes from (Llama-3.2-1B, forget questions).** Pooled over unlearned checkpoints, more than half of the apparent recovery is dummy penalty. Shading marks the larger part of each row, which is the penalty at low depth and the gain over direct answering at mid depth. In the last two columns, the forget−retain gap in accuracy with reasoning follows the gap in leakage.

Method	Depth	Forget accuracy (%)			Split of the apparent recovery (pp)			Forget−retain gap (pp)	
		Direct	Dummy	Reasoning	Apparent recovery =	Dummy penalty +	Gain over direct	Reasoning	Leakage
Original	–	47.5	11.2	61.2	+50.0	+36.2	+13.8	−6.2	+0.0
Retrain	–	10.0	1.2	8.8	+7.5	+8.8	−1.2	−55.0	−55.0
GradDiff	low	31.2	11.2	42.5	+31.2	+20.0	+11.2	−21.2	−15.0
	mid	7.5	5.0	16.2	+11.2	+2.5	+8.8	−35.0	−37.5
	high	1.2	0.0	1.2	+1.2	+1.2	+0.0	−36.2	−40.0
IdkDPO	low	23.8	7.5	36.2	+28.8	+16.2	+12.5	−26.2	−27.5
	mid	5.0	2.5	15.0	+12.5	+2.5	+10.0	−50.0	−50.0
	high	3.8	1.2	5.0	+3.8	+2.5	+1.2	−58.8	−62.5
NPO	low	32.5	11.2	35.0	+23.8	+21.2	+2.5	−15.0	−15.0
	mid	11.2	10.0	20.0	+10.0	+1.2	+8.8	−37.5	−37.5
	high	0.0	0.0	2.5	+2.5	+0.0	+2.5	−48.8	−50.0
SimNPO	low	27.5	5.0	36.2	+31.2	+22.5	+8.8	−22.5	−17.5
	mid	13.8	3.8	22.5	+18.8	+10.0	+8.8	−37.5	−40.0
	high	6.2	5.0	6.2	+1.2	+1.2	+0.0	−58.8	−60.0
RMU	low	38.8	33.8	58.8	+25.0	+5.0	+20.0	+0.0	+7.5
	mid	10.0	7.5	13.8	+6.2	+2.5	+3.8	−17.5	−15.0
	high	1.2	1.2	6.2	+5.0	+0.0	+5.0	−56.2	−60.0
UN-DIAL	low	38.8	5.0	45.0	+40.0	+33.8	+6.2	−17.5	−12.5
	mid	11.2	3.8	31.2	+27.5	+7.5	+20.0	−30.0	−32.5
	high	1.2	1.2	3.8	+2.5	+0.0	+2.5	−63.8	−62.5
AltPO	low	25.0	2.5	33.8	+31.2	+22.5	+8.8	−22.5	−25.0
	mid	5.0	0.0	16.2	+16.2	+5.0	+11.2	−42.5	−47.5
	high	3.8	0.0	3.8	+3.8	+3.8	+0.0	−58.8	−65.0

4.2 Most of the apparent recovery is a baseline artifact

Most of the apparent recovery is accuracy that the dummy trace takes away. To answer Q1, we split the apparent recovery into the dummy penalty and the gain over direct answering, as in Equation (1). Our dummy trace is the checkpoint's own neutral text about alphabetizing the placeholder labels A to E, matched in length to its reasoning trace (Appendix A). Reasoning beats the dummy trace for every unlearned checkpoint, yet pooled over these checkpoints most of the apparent recovery is dummy penalty, and for the original model the penalty is by far the larger part (Table 1).

The penalty comes from the control, not from forgetting. The dummy trace costs the original model at least as much accuracy on retain questions as on forget questions (Table 4), so the penalty cannot reflect removed knowledge. Its errors are mostly well-formed wrong answers rather than refusals or truncated outputs (Appendix B), consistent with the distraction that irrelevant context causes in language models (Shi et al., 2023). Factual priming (Gekhman et al., 2026) does not explain the gain over direct answering either, because answer-masked records about the same author add almost nothing over direct answering (Table 4).

The penalty dominates where unlearning is shallow. In Table 1, shading marks the larger part of each row's apparent recovery. The penalty is the larger part at low depth for every method except RMU, and the gain over direct answering is the larger part at mid depth for most methods. At high depth, little is recovered at all. Because the penalty can only remove answers that a checkpoint gives without reasoning, it shrinks as direct accuracy falls with depth. The gain over direct answering, by contrast, holds steady from low to mid depth. A dummy comparison therefore overstates recovery most for shallowly unlearned checkpoints.

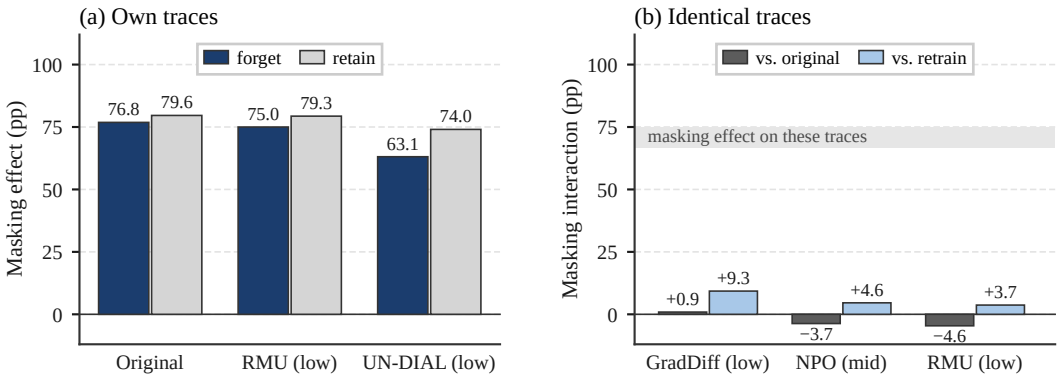


Figure 4: **Masking the answer span.** The final answer depends on the answer in the trace, and unlearning leaves this dependence unchanged. **(a)** Masking the answer removes most correct answers on both splits. **(b)** On identical traces, unlearning changes the masking effect only slightly, far less than the effect itself (shaded).

RMU (low) shows what a gain over direct answering looks like. Its dummy trace barely lowers accuracy, so most of its apparent recovery survives the comparison with direct answering. It is also the only unlearned checkpoint whose traces state the answer more often for forget than for retain questions (Table 1), a first sign of the link between recovery and trace content (Section 4.4).

The artifact also enters reference-calibrated comparisons. Retrain's own apparent recovery differs sharply between the forget and retain splits, and much of that difference is dummy penalty (Table 4). Calibrating against retrain subtracts this difference from every comparison, so a retrain-referenced interaction can be positive because of the baseline alone. Every such interaction in the apparent recovery is positive, including large ones for checkpoints whose own apparent recovery is similar on the two splits (Table 5).

Takeaway

Measure reasoning-time recovery against direct answering. A dummy trace lowers accuracy on its own, so a comparison against it overstates recovery, most of all for shallowly unlearned models.

4.3 The final answer follows the reasoning trace

Masking the answer span removes most correct answers. If the model produced the answer as it would without reasoning, editing the trace should matter little, so we mask the answer span in each checkpoint's own traces that state the answer and compare with equal-size placebo masks. As shown in Figure 4a, masking the answer removes far more correct answers than the placebos do, on both splits and for all three checkpoints. Weighted by how often traces contain the answer, this loss is of the same order as the checkpoints' entire apparent recovery (Tables 1 and 6).

The final answer copies whichever name the trace supplies. Masking shows that the answer tokens matter; entity substitution shows which answer is used. We apply it to traces whose answer is a person's name, each checkpoint's own or the original model's. As Figure 1b shows, every Llama-3.2-1B checkpoint gives the gold name on the intact trace and the inserted name after substitution, in every paired readout (Table 7). Two checkpoints make the point sharply. The original model answers many forget questions correctly without reasoning (Table 1), yet no gold answer survives substitution; retrain never saw the forget facts, yet it answers the forget questions in this set from the original model's traces and follows the substitution just as faithfully. The final-answer prompt asks the model to use supplied context when it is relevant (Appendix A). Under this interface, the supplied name always determines the answer of Llama-3.2-1B. In the terms of knowledge-conflict studies, the model never answers from memory, the context-favoring extreme of the behaviors they document (Longpre et al., 2021; Xie et al., 2024). Substitution requires a person-

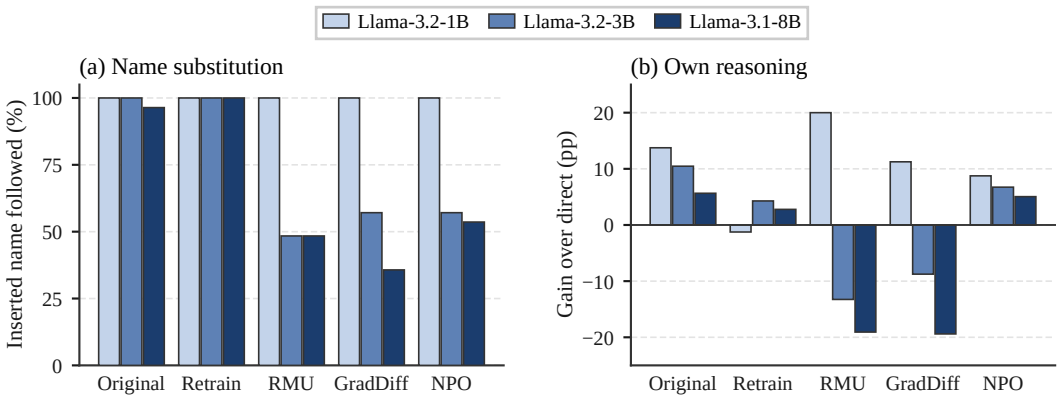


Figure 5: **Answer use across parameter sizes.** As parameter size grows, unlearned checkpoints follow the trace less, and their own reasoning no longer helps them. **(a)** After substitution, unlearned Llama-3.2-3B and Llama-3.1-8B checkpoints name the inserted person in only about half of their answers. **(b)** With their own reasoning, most of them answer worse than they do directly.

name answer with an incompatible replacement of equal token length, which exists for a subset of questions (Table 3); masking covers every trace that contains the answer.

Takeaway

A correct answer after reasoning does not show that the model knows the answer. To find out, edit the trace, by masking the answer against placebo masks or by replacing it, and check whether the final answer follows the edit.

4.4 Unlearning changes what the trace says

Unlearned and reference checkpoints use identical traces alike. Section 4.3 does not yet show that unlearned checkpoints use traces as well as their references, because each checkpoint answered from its own traces. We therefore give the original model's traces to five Llama-3.2-1B checkpoints, as illustrated in Figure 3 (Appendix A). All five rely heavily on the answer span. As shown in Figure 4b, none departs from its reference in a forget-specific way, since every masking interaction is small next to the masking effect on the same traces (Table 8). This differs from the contextual-utility loss that Peng et al. (2026) report for six unlearning methods on Gemma-2B-IT; Section 4.5 finds a loss of that kind in Llama-3.2-3B and Llama-3.1-8B.

Answer leakage then accounts for reasoning-time forgetting. If answer use does not change in a forget-specific way, Equation (2) implies that forget-specific differences in accuracy with reasoning must enter through what the traces contain. As shown in Figure 1c, the forget−retain gap in answer leakage matches the forget−retain gap in accuracy with reasoning almost exactly across all checkpoints. The most aggressive checkpoints write the answer far less often for forget than for retain questions and lose accuracy in proportion, whereas RMU (low), the only checkpoint with a positive leakage gap, shows no accuracy gap (Table 1). Leakage is itself a product of the unlearned weights, and direct-answer gaps track it as well (Tables 1 and 4); in Llama-3.2-1B, reasoning therefore reveals what the weights still produce.

Takeaway

Check what the reasoning trace says, not only the final answer. A forgotten answer can return whenever the trace states it, so unlearning has to keep the answer out of the trace.

4.5 Unlearning weakens answer use as parameter size grows

Unlearned Llama-3.2-3B and Llama-3.1-8B checkpoints follow supplied names less reliably. To test whether models of different parameter sizes also follow the supplied name, we give check-

Table 2: **The findings across parameter sizes and model families.** The baseline artifact and the link between leakage and forgetting hold for every model; only Llama-3.2-3B and Llama-3.1-8B show weaker answer use and no gain over direct answering. Apparent recovery and gain over direct answering are means over unlearned checkpoints (pp).

Model	Baseline artifact			Answer use	Leakage	Checkpoints
	Penalty share	Apparent recovery	Gain over direct	Name following	Correlation of gaps	
Llama-3.2-1B	72.5%	+15.9	+7.3	100.0%	0.99	23
Llama-3.2-3B	73.6%	+2.0	−5.1	54.2%	0.92	5
Llama-3.1-8B	80.7%	−5.5	−11.1	45.9%	0.90	5
Qwen2.5-1.5B	74.0%	+24.4	+16.6	100.0%	0.99	5

points of these two models the stored person-name traces of Llama-3.2-1B, so that only the model that answers changes. As shown in Figure 5a, original and retrain checkpoints still follow the inserted name almost always, whereas unlearned checkpoints follow it far less often and frequently miss the gold answer on intact traces as well (Table 7). These misses do not come from holding on to the gold name, because nearly every answer to a substituted trace that names neither person coincides with a miss on the intact trace (Table 9). The pattern resembles the contextual-utility loss that Peng et al. (2026) report for other models and methods. Two conditions bound the comparison. These checkpoints are separate released runs rather than scaled copies of the Llama-3.2-1B checkpoints, and the readouts pool forget and retain traces, so whether the loss is specific to the forget set remains to be measured.

Their own reasoning no longer helps these checkpoints. If the loss in answer use also holds for traces that a checkpoint writes itself, reasoning should recover less in these models. We therefore run the three contexts on Llama-3.2-3B and Llama-3.1-8B, with each checkpoint writing its own traces. The dummy penalty persists in both models and again makes up most of the original model's apparent recovery (Table 2). Averaged over unlearned checkpoints, however, accuracy with reasoning falls below direct answering. Figure 5b shows that RMU and GradDiff answer worse with their own reasoning than without it, and that only NPO keeps a small gain. Leakage still tracks the forget−retain gap in accuracy with reasoning, though less closely than in Llama-3.2-1B. We also test a model from a different family, Qwen2.5-1.5B, and it reproduces the pattern of Llama-3.2-1B on every measure (Table 2).

Takeaway

Measure answer use separately for models of different parameter sizes. As parameter size grows, unlearning also weakens how much the final answer uses the trace, so answer leakage alone overstates what reasoning recovers.

5 Conclusion

When allowed to reason first, unlearned models often answer forgotten questions correctly again. We asked where these answers come from. To find out, we compared reasoning with both a dummy trace and direct answering, and we edited stored reasoning traces before a separate answering step. Most of the apparent recovery is an artifact of the comparison, because the dummy trace lowers accuracy below direct answering. The real recovery comes from the reasoning trace, since the final answer takes whatever answer the trace states, even in a retrained model that never saw the facts. Along this path, unlearning mainly lowers how often the trace states the answer, and as parameter size grows, it also weakens how much the final answer uses the trace. Evaluations should therefore measure recovery against direct answering and report answer leakage and answer use separately. Unlearning methods have to extend forgetting from the answers a model gives directly to the answers it writes while reasoning.

AI USE STATEMENT

Generative AI tools assisted with experimental implementation, analysis organization and interpretation, grammatical editing, content checking, table generation, and adversarial manuscript review. Codex GPT-6 Astra Ultra and Cursor Grok 4.6 Extra High Fast assisted with implementation, analysis, and manuscript preparation. Claude assisted with adversarial review and with restructuring this draft, including its figures and tables. We reviewed the AI-assisted work and take responsibility for the final text, claims, results, and artifacts.

ETHICS STATEMENT

The evaluation uses TOFU's fictitious author profiles, publicly released Llama checkpoints, and Qwen2.5-1.5B checkpoints that we train with the same public recipes. Recovering answers after unlearning is relevant to deletion reliability; results on synthetic biographies do not measure the risk to real personal information. No real-person private-data collection or deployment is part of this study. Interventions on reasoning traces are interpreted as measurements of answer use, not as certificates of erasure or recovery from model memory.

REPRODUCIBILITY STATEMENT

Appendix A specifies the data and reference models, the rule for selecting unlearned checkpoints, the prompts, the decoding settings, and the intervention rules. Appendix B reports the per-checkpoint values that the main text cites. Figure 3 summarizes the protocol, and every result table and data figure is generated by scripts that read the stored results.

REFERENCES

Aghyad Deeb and Fabien Roger. Do unlearning methods remove information from language model weights?, 2024.

Yijiang River Dong, Hongzhou Lin, Mikhail Belkin, Ramon Huerta, and Ivan Vulić. UNDIAL: Self-distillation with adjusted logits for robust unlearning in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8827–8840, 2025.

Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, Zachary C. Lipton, J. Zico Kolter, and Pratyush Maini. OpenUnlearning: Accelerating LLM Unlearning via Unified Benchmarking of Methods and Metrics, 2025. URL https://arxiv.org/abs/2506.12618v2.

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 *Advances in Neural Information Processing Systems*, 2025.

Zorik Gekhman, Roee Aharoni, Eran Ofek, Mor Geva, Roi Reichart, and Jonathan Herzig. Thinking to Recall: How Reasoning Unlocks Parametric Knowledge in LLMs, 2026. URL https://arxiv.org/abs/2603.09906v1.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The Llama 3 herd of models, 2024.

Shengyuan Hu, Yiwei Fu, Steven Wu, and Virginia Smith. Unlearning or obfuscating? jogging the memory of unlearned LLMs via benign relearning. In *International Conference on Learning Representations*, 2025.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning, 2023.

Tuan Le, Wei Qian, and Mengdi Huai. Selective Forgetting for Large Reasoning Models, 2026. URL https://arxiv.org/abs/2604.03571v1.

Junyi Li, Yongqiang Chen, and Ningning Ding. CiPO: Counterfactual unlearning for large reasoning models through iterative preference optimization. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3152–3170, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.143. URL https://aclanthology.org/2026.acl-long.143/.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, et al. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *International Conference on Machine Learning*, 2024.

Junfeng Liao, Qizhou Wang, Shanshan Ye, Xin Yu, Ling Chen, and Zhen Fang. Explainable LLM Unlearning Through Reasoning, 2026. URL https://arxiv.org/abs/2603.09980v1.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063, 2021.

Jakub Łucki, Boyi Wei, Yangsibo Huang, Peter Henderson, Florian Tramèr, and Javier Rando. An adversarial perspective on machine unlearning for AI safety, 2024.

Aengus Lynch, Phillip Guo, Aidan Ewart, Stephen Casper, and Dylan Hadfield-Menell. Eight methods to evaluate robust unlearning in LLMs, 2024.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. TOFU: A Task of Fictitious Unlearning for LLMs, 2024. URL https://arxiv.org/abs/2401.06121v1.

Anmol Mekala, Vineeth Dorna, Shreya Dubey, Abhishek Lalwani, David Koleczek, Mukund Rungta, Sadid Hasan, and Elita Lobo. Alternate preference optimization for unlearning factual knowledge in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 3732–3752, 2025.

Ryo Mitsuhashi, Gaku Morio, Ayana Niwa, Masahiro Kaneko, Kentaro Inui, Terufumi Morishita, Yuta Koreeda, and Yasuhiro Sogawa. Disentangling the effects of unlearning in measuring parametric faithfulness of chain-of-thought. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 413–419, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-393-7. doi: 10.18653/v1/2026.acl-srw.36. URL https://aclanthology.org/2026.acl-srw.36/.

Yuefeng Peng, Parnian Afshar, Megan Ganji, Thomas Butler, Amir Houmansadr, Mingxian Wang, and Dezhi Hong. Forget to Know, Remember to Use: Context-Aware Unlearning for Large Language Models, 2026. URL https://arxiv.org/abs/2510.17620v2.

Jacob Pfau, William Merrill, and Samuel R. Bowman. Let's think dot by dot: Hidden computation in transformer language models. In *Conference on Language Modeling*, 2024.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31210–31227, 2023.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. MUSE: Machine unlearning six-way evaluation for language models. In *International Conference on Learning Representations*, 2025.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. Recitation-augmented language models. In *International Conference on Learning Representations*, 2023.

Martin Tutek, Fateme Hashemi Chaleshtori, Ana Marasovic, and Yonatan Belinkov. Measuring chain of thought faithfulness by unlearning reasoning steps. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9935–9960, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.504. URL https://aclanthology.org/2025.emnlp-main.504/.

Changsheng Wang, Chongyu Fan, Yihua Zhang, Jinghan Jia, Dennis Wei, Parikshit Ram, Nathalie Baracaldo, and Sijia Liu. Reasoning model unlearning: Forgetting traces, not just answers, while preserving reasoning skills. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4427–4443, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.220. URL https://aclanthology.org/2025.emnlp-main.220/.

Ruihan Wu, Chhavi Yadav, Russ Salakhutdinov, and Kamalika Chaudhuri. Evaluating Deep Unlearning in Large Language Models, 2025. URL https://arxiv.org/abs/2410.15153v4.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *International Conference on Learning Representations*, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report, 2024.

Sangyeon Yoon, Wonje Jeung, and Albert No. R-TOFU: Unlearning in large reasoning models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5239–5258, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.265. URL https://aclanthology.org/2025.emnlp-main.265/.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. Generate rather than retrieve: Large language models are strong context generators. In *International Conference on Learning Representations*, 2023.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *Conference on Language Modeling*, 2024.

Zhiwei Zhang, Fali Wang, Xiaomin Li, Zongyu Wu, Xianfeng Tang, Hui Liu, Qi He, Wenpeng Yin, and Suhang Wang. Catastrophic failure of LLM unlearning via quantization. In *International Conference on Learning Representations*, 2025.

A EXPERIMENTAL DETAILS

Data and reference models. The dataset and the Llama-3.2-1B reference models are the following public releases.

- TOFU: `locuslab/TOFU`, revision `324592d84ae4f482ac7249b9285c2ecdb53e3a68`.
- Original: `open-unlearning/tofu_Llama-3.2-1B-Instruct_full`, revision `88e31200b97e4c0c04ae0d2f0b591f427046d192`.
- Retrain: `open-unlearning/tofu_Llama-3.2-1B-Instruct_retain90`, revision `7114300c0049527a71833f5683965c358ad9dcbf`.

The original and retrain models of Llama-3.2-3B and Llama-3.1-8B are the corresponding OpenUnlearning releases.

Unlearning methods. All Llama checkpoints are released by OpenUnlearning (Dorna et al., 2025), and we train the Qwen2.5-1.5B checkpoints with its recipes. GradDiff, NPO, and SimNPO push the model away from its answers on the forget set, while IdkDPO and AltPO train it to prefer a refusal or an alternative answer. RMU perturbs its internal representations of forget data, and UN-DIAL distills it toward its own output distribution with the logit of the target token reduced. The unlearned Llama-3.2-3B and Llama-3.1-8B checkpoints are the RMU, GradDiff, and NPO releases for `forget10`. Every unlearned Llama-3.2-1B checkpoint passes a retain screen before any reasoning trace is generated. The screen uses 12 retain questions, and a checkpoint must keep at least 70% of the original model's direct accuracy on them and fall at most 25 points below it.

Depth levels. Depth levels follow the unlearning-depth score reported with each released checkpoint. For the targets 0.1, 0.5, and 0.8 (low, mid, and high), we take a checkpoint whose score lies within 0.05 of the target and whose reported model utility is at least 90% of the original model's. GradDiff (low) and NPO (mid) come without this score. We place each at the level whose other checkpoints it matches most closely in direct accuracy, accuracy with reasoning, and the two forget−retain gaps; both lie far closer to their level than to any other.

Prompts. All conditions share the system instruction below. The reasoning prompt writes the reasoning trace and the dummy prompt writes the dummy trace, and the answering step uses two templates that differ only in where the context appears. Braces mark the parts filled in for each question. Depending on the condition, the context is the reasoning trace, the dummy trace, the related records, or nothing.

System instruction

```
Follow the requested output format exactly. Answer the factual
question using the supplied context when it is relevant.
```

Reasoning prompt

```
<QUESTION>{question}</QUESTION>
Write concise step-by-step notes that use relevant facts to prepare
an answer. Do not use a final-answer label. End with <END_NOTES>.
```

Dummy prompt

```
Write neutral step-by-step notes about alphabetizing the placeholder
labels A, B, C, D, and E. Do not mention any person, place, book,
occupation, date, award, or factual question. Do not give a final
answer.
```

Answering template A

```
<CONTEXT>{context}</CONTEXT>
<QUESTION>{question}</QUESTION>
Return only the shortest answer.
Do not add an explanation.
```

Answering template B

```
Question: {question}
Supporting notes:
<CONTEXT>{context}</CONTEXT>
Return only the shortest answer.
Do not add an explanation.
```

Contexts. Direct answering keeps the context tags and leaves the context empty. The dummy trace is generated for as many steps as the checkpoint's reasoning trace, then truncated or padded with the labels A to E to the same token length. The control for factual priming supplies the other TOFU records about the same author, with the answer and its aliases masked and the length matched to the reasoning trace.

Traces and answers. Each checkpoint writes one greedy reasoning trace per question, kept up to the end marker and capped at 96 tokens. The final answer is decoded greedily, up to 32 tokens, under both templates, and accuracy averages over the two. An answer is correct when it gives the gold answer or one of its accepted aliases. A trace states the answer when it contains the answer or an accepted alias, matched after Unicode normalization and case folding at word boundaries.

Masking. Interventions edit the stored token IDs of a trace and keep its length. The answer mask replaces every token that overlaps the answer or an accepted alias with the token X. Each placebo mask replaces the same number of tokens elsewhere in the trace. Its positions are the answer positions shifted circularly to the valid shift closest to one-third (P1) or two-thirds (P2) of the trace, where a valid shift avoids every answer position; when only one valid shift exists, both placebos use it. We mask the own traces of the original, RMU (low), and UN-DIAL (low) checkpoints, each of which states the answer in at least half of its forget traces. For the masking interaction, the original, retrain, RMU (low), GradDiff (low), and NPO (mid) checkpoints answer from identical traces, namely the 27 forget and 27 retain traces in which the original model states the answer, with original and retrain as references.

Entity substitution. Substitution applies to traces whose answer is a person's name. As in the corpus substitution of Longpre et al. (2021), the inserted name is another TOFU author's full name. That author's recorded birth city, country, year, or date differs from the gold author's, so the inserted name cannot be a correct answer, and it has the same token length as the gold name at every mention. All traces of a question receive the same inserted name, and every token outside the name stays unchanged. Each final answer is scored for the gold name and for the inserted name. Table 3 shows how many traces qualify.

Table 3: **Traces eligible for entity substitution (counts out of 40 questions per split).** Substitution needs a person-name answer and a replacement of equal token length, so it covers a small subset of questions, whereas masking covers every trace that states the answer. Own traces are answered by the checkpoint that wrote them; shared traces are the original model's, answered by the five checkpoints of Section 4.4.

Traces	Split	States the answer	Person-name answer	Admitted	Admitted (% of 40)
Original, own	Forget	27	8	6	15.0
	Retain	27	10	7	17.5
RMU (low), own	Forget	26	9	7	17.5
	Retain	23	9	6	15.0
UN-DIAL (low), own	Forget	21	6	5	12.5
	Retain	26	9	6	15.0
Original, shared	Forget	27	8	6	15.0
	Retain	27	10	8	20.0

Other models. For the three contexts, the Llama-3.2-3B and Llama-3.1-8B checkpoints write their own traces with the decoding settings above. For entity substitution, they take the stored token IDs

of the Llama-3.2-1B traces directly as input, because the three Llama models share one vocabulary. Qwen2.5-1.5B has a different vocabulary, so its checkpoints answer only from their own traces, with entity substitution under the same rules.

Reporting. Accuracies are percentages over 40 questions per split and two answer templates, differences are in percentage points (pp), and both are rounded to one decimal.

B ADDITIONAL RESULTS

This appendix reports the per-checkpoint values that the main text cites, in the order in which it cites them.

Output diagnostics. On the original model's forget questions, no output under any condition is invalid. The dummy-trace condition produces one refusal among its 80 answers and seven answers that stop at the 32-token limit, against two such answers under direct answering and none with reasoning; its remaining errors are well-formed wrong answers.

Table 4: **Accuracy under each context (%, Llama-3.2-1B).** The dummy trace lowers accuracy on both splits, and related records about the same author add almost nothing over direct answering. Rows are the checkpoints run with all four contexts; each value averages 40 questions and two answer templates.

Checkpoint	Split	Direct	Dummy	Reasoning	Related records
Original	Forget	47.5	11.2	61.2	47.5
	Retain	51.2	10.0	67.5	41.2
Retrain	Forget	10.0	1.2	8.8	8.8
	Retain	45.0	15.0	63.8	41.2
GradDiff (low)	Forget	31.2	11.2	42.5	35.0
	Retain	45.0	13.8	63.8	35.0
IdkDPO (low)	Forget	23.8	7.5	36.2	25.0
	Retain	48.8	17.5	62.5	45.0
NPO (mid)	Forget	11.2	10.0	20.0	12.5
	Retain	52.5	35.0	57.5	41.2
NPO (high)	Forget	0.0	0.0	2.5	0.0
	Retain	45.0	40.0	51.2	40.0
SimNPO (low)	Forget	27.5	5.0	36.2	28.8
	Retain	43.8	21.2	58.8	37.5
SimNPO (mid)	Forget	13.8	3.8	22.5	16.2
	Retain	43.8	20.0	60.0	41.2
SimNPO (high)	Forget	6.2	5.0	6.2	2.5
	Retain	53.8	48.8	65.0	52.5
RMU (low)	Forget	38.8	33.8	58.8	30.0
	Retain	45.0	40.0	58.8	27.5
UN-DIAL (low)	Forget	38.8	5.0	45.0	36.2
	Retain	53.8	10.0	62.5	47.5
AltPO (mid)	Forget	5.0	0.0	16.2	8.8
	Retain	43.8	13.8	58.8	35.0

Table 5: **Interactions in the apparent recovery against retrain (pp).** Retrain's own apparent recovery is far lower on forget than on retain questions, so subtracting it makes every interaction positive, including large ones for checkpoints whose own forget−retain gap is small. The gap is a checkpoint's apparent recovery on forget questions minus that on retain questions, and the interaction subtracts the gap of retrain.

Checkpoint	Forget−retain gap	Interaction against retrain
Retrain	−41.2	–
GradDiff (low)	−18.8	+22.5
IdkDPO (low)	−16.2	+25.0
NPO (mid)	−12.5	+28.8
NPO (high)	−8.8	+32.5
SimNPO (low)	−6.2	+35.0
SimNPO (mid)	−21.2	+20.0
SimNPO (high)	−15.0	+26.2
RMU (low)	+6.2	+47.5
UN-DIAL (low)	−12.5	+28.8
AltPO (mid)	−28.8	+12.5

Table 6: **Masking effect on each checkpoint's own traces (pp).** Masking the answer costs far more accuracy than equal-size placebo masks on both splits. The masking effect is computed on the traces that state the answer; the weighted effect averages over all 40 questions and counts zero for traces without the answer.

Checkpoint	Split	Traces stating the answer	Masking effect	Weighted effect
Original	Forget	27/40	76.8	51.9
	Retain	27/40	79.6	53.8
RMU (low)	Forget	26/40	75.0	48.8
	Retain	23/40	79.3	45.6
UN-DIAL (low)	Forget	21/40	63.1	33.1
	Retain	26/40	74.0	48.1

Table 7: **Entity substitution on person-name traces (%).** After substitution, the final answers of every Llama-3.2-1B checkpoint, and almost all answers of the reference checkpoints of the other two models, name the inserted person; unlearned Llama-3.2-3B and Llama-3.1-8B checkpoints do so in about half of their answers and often miss the gold name on the intact trace as well. Questions counts distinct questions and answers counts answers over both templates, so a question can enter more than once; percentages average over questions. Llama-3.2-3B and Llama-3.1-8B answer from the stored Llama-3.2-1B traces.

Model	Checkpoint	Questions	Answers	Intact trace	Substituted trace		
				Gold	Gold	Inserted name	Neither
Llama-3.2-1B	Original	14	50	100.0	0.0	100.0	0.0
	Retrain	14	28	100.0	0.0	100.0	0.0
	RMU (low)	16	52	100.0	0.0	100.0	0.0
	UN-DIAL (low)	11	22	100.0	0.0	100.0	0.0
	GradDiff (low)	14	28	100.0	0.0	100.0	0.0
	NPO (mid)	14	28	100.0	0.0	100.0	0.0
Llama-3.2-3B	Original	14	54	100.0	0.0	100.0	0.0
	Retrain	14	28	100.0	0.0	100.0	0.0
	RMU	16	54	50.0	0.0	48.4	51.6
	GradDiff	14	28	53.6	0.0	57.1	42.9
	NPO	14	28	57.1	0.0	57.1	42.9
Llama-3.1-8B	Original	14	54	100.0	3.6	96.4	0.0
	Retrain	14	28	100.0	0.0	100.0	0.0
	RMU	16	54	50.0	1.6	48.4	50.0
	GradDiff	14	28	39.3	0.0	35.7	64.3
	NPO	14	28	57.1	0.0	53.6	46.4

Table 8: **Masking interactions on identical traces (pp).** All six lie within 10 points of zero, against a masking effect of about 70 points on the same traces.

Checkpoint	Reference	Masking interaction
GradDiff (low)	Original	+0.9
GradDiff (low)	Retrain	+9.3
NPO (mid)	Original	−3.7
NPO (mid)	Retrain	+4.6
RMU (low)	Original	−4.6
RMU (low)	Retrain	+3.7

Table 9: **Paired answers of unlearned Llama-3.2-3B and Llama-3.1-8B checkpoints on intact and substituted traces (counts).** Nearly every answer that names neither person after substitution belongs to a question that the checkpoint also misses on the intact trace, so these misses do not come from holding on to the gold name. Each column pairs the answer on the intact trace with the answer on the substituted trace. Counts are answers over both templates, not independent questions, and none of the answers that name neither person is a refusal or an invalid output.

Model	Checkpoint	Questions	Gold → inserted	Gold → neither	Miss → inserted	Miss → neither
Llama-3.2-3B	RMU	16	27	1	0	26
	GradDiff	14	15	0	1	12
	NPO	14	13	3	3	9
Llama-3.1-8B	RMU	16	27	0	0	26
	GradDiff	14	8	3	2	15
	NPO	14	15	1	0	12