Unlearning Is Not Just Erasing: Temporal Decoupling via Generation Inequality

Xunlei Chen[1] Qirui Ye[1], Yuang Li[1], Yi Gong[1],
Zhaokun Wang[1], Wenyi Li[1], Shiyao Guo[1], Jinyu Guo[2]

[1]School of Information and Software Engineering, University of Electronic Science and Technology of China
[2]School of Computer Science and Engineering, University of Electronic Science and Technology of China
xunlei@std.uestc.edu.cn

Abstract

Large language models (LLMs) require effective unlearning to address privacy regulations and safety concerns. However, achieving precise forgetting without compromising general utility remains challenging. Existing sequence- and token-level methods penalize target outputs without modeling their context-dependent retrieval paths, which can disrupt linguistic structure or suppress benign knowledge. We present **ADU**, a fine-grained, training-based framework that shifts unlearning from token erasure to contextual attention-pathway decoupling. Exploiting the functional distinction between local and global attention heads, ADU identifies preplan positions that retrieve persistent sensitive anchors and fixes their candidate paths under the original model. It then trains attention-projection adapters to suppress attention mass along these paths while preserving local-attention structure and retain-set language modeling. Post-training activation exchange tests whether the modified attention-output module transmits the learned forgetting effect. ADU achieves the strongest aggregate performance among evaluated baselines on the TOFU and WMDP benchmarks, including a Forget Quality of 0.93 on TOFU. It preserves 87–98% of model utility (92.9% on average versus 81.9% for baselines) while reducing side effects in benign contexts.

Introduction

Large language models (LLMs) have achieved advances in language understanding and generation (Lee et al. 2020; Ranjan et al. 2026) through model scaling (Hu et al. 2025) and diverse pretraining data (Zhao et al. 2025a). However, models can reproduce sensitive content (Gong et al. 2026), private information (Maini et al. 2024), or unsafe material (Shi et al. 2024a, 2025). As the *General Data Protection Regulation* (GDPR) and the *"right to be forgotten"* receive attention (Grynbaum et al. 2023), machine unlearning has emerged as a privacy mechanism (Zhao et al. 2025b). It aims to reduce the accessibility of requested knowledge while preserving unrelated model capabilities (Yu et al. 2025).

Balancing unlearning efficacy and model utility remains fundamentally challenging (Pu et al. 2026; Cha et al. 2024). Prompt-based (Bhaila, Van, and Wu 2025; Wang et al. 2026) and auxiliary-model approaches (Geng et al. 2025; Sun et al. 2025) can change outputs without updating the target model, yet knowledge may remain recoverable through extraction attacks (Shah et al. 2025). Parameter-level unlearning remains necessary when the objective is to modify the deployed model itself, particularly for private or copyrighted content.

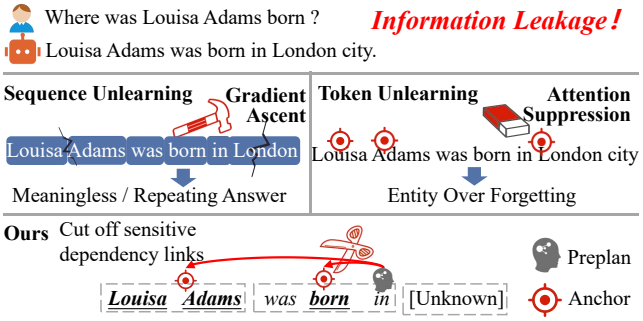
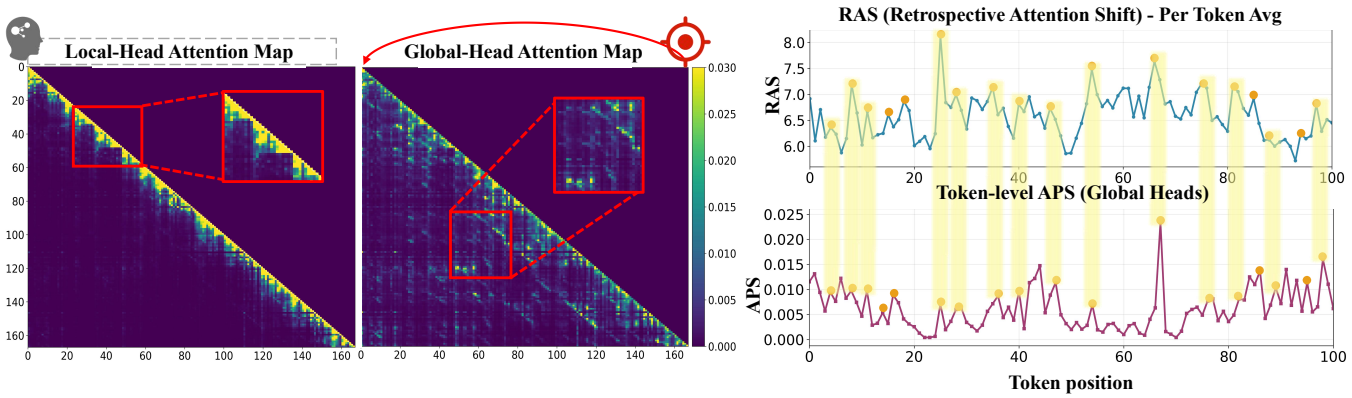


Figure 1: Sequence methods minimize the joint probability of the entire QA pair. Token methods blindly suppress the probabilities of individual tokens. We sever attention pathways that point to the anchor tokens.

Existing parameter-level methods are categorized by their optimization target (Zhuang et al. 2025; Kim et al. 2026). Sequence-level unlearning treats an entire question–answer sequence as the forget target (Liu et al. 2025). Because its loss spans syntactic and content tokens, it can damage linguistic structure and generation fluency (Nguyen et al. 2025). Token-level unlearning instead penalizes selected sensitive tokens (Jiang et al. 2025) and limits damage by avoiding noncritical positions (Yu et al. 2025; Lee, Liu, and Xiong 2026). Nevertheless, both approaches define forgetting over visible output targets rather than the internal, context-dependent computation through which sensitive knowledge is retrieved.

This distinction matters because the same entity can be sensitive in one context and benign in another. Static sequence or token penalties may therefore suppress legitimate occurrences, causing excessive forgetting and degraded factual behavior in non-sensitive contexts (Tan et al. 2025). As illustrated in Fig. 1, the desired intervention should target the corresponding retrieval computation rather than erase the entity itself. This raises a natural question: *Can an LLM forget a context-specific retrieval route without destabilizing ordinary generation?*

Motivated by evidence that attention heads contribute unevenly to model behavior (Lin et al. 2025), we call their asym-



(a) Visualization of local and global attention patterns on MMLU

(b) Token-level RAS and APS trajectories on a WMDP sample

Figure 2: Generation inequality and temporal retrieval rhythm. (a) Local heads preserve near-diagonal dependencies, global heads form long-range patterns. (b) RAS peaks indicate preplan transitions, APS peaks identify persistent anchors, shaded bands mark selected regions. Post-training edge-contribution replacement tests path-specific causal mediation after unlearning.

metric temporal roles *generation inequality*. Local heads maintain short-range dependencies, whereas global heads route information from distant, persistently attended tokens. Around semantic transitions, local attention shifts retrospectively at preplan positions, after which global heads retrieve earlier anchors that shape subsequent generation. Average Backward Distance separates the two head groups, while Retrospective Attention Shift and Anchor Persistence Score locate candidate preplans and anchors (Fig. 2).

Based on this rhythm, we propose **A**ttention **D**ecoupling **U**nlearning (**ADU**), a training-based method for context-specific pathway suppression. ADU computes the head partition and preplan–anchor paths once under the original model and keeps them fixed during training. Attention-projection adapters reduce path mass, while retained language modeling and local-attention preservation constrain collateral changes. The learned decoupling operates in every forward pass; bidirectional edge-contribution replacement runs only after training to test whether the selected preplan–anchor paths causally mediate the resulting forgetting effect.

Unlike representation-level methods such as RMU, token-editing methods such as MET, and broad attention suppression such as ASU, ADU optimizes a context-indexed retrieval pathway rather than a hidden representation or token identity. Attention weights locate candidate routes but are not treated as explanations by themselves: the pathway objective controls transported contributions under bounded values, while bidirectional edge-contribution replacement provides post-training tests of path-specific causal mediation. Across TOFU, WMDP, and MUSE-Harry Potter, ADU achieves the strongest aggregate forgetting–retention trade-off among the evaluated baselines. It obtains 93% Forget Quality on TOFU and preserves 87–98% of model utility, averaging 92.9% compared with 81.9% for baselines. These results support contextual pathway suppression as a more targeted alternative to sequence or token erasure. In summary, our contributions are:

1. We formulate LLM unlearning as contextual pathway decoupling and characterize a preplan–anchor temporal pattern arising from the functional specialization of local and global attention heads.

2. We propose ADU, which trains attention-projection adapters to suppress fixed preplan-to-anchor pathways while retaining language modeling and local-attention preservation constrain collateral behavior.

3. We provide conditional theoretical guarantees and extensive evaluations on WMDP, TOFU, and MUSE-Books, demonstrating state-of-the-art forgetting retention trade-offs and validating the causal role of the targeted pathways.

Related Work

Sequence-level Unlearning Many parameter-level methods optimize a coarse sequence-level objective. Gradient Ascent (GA) (Thudi et al. 2022) directly maximizes the forget-set loss and can produce unstable parameter updates. Negative Preference Optimization (NPO) (Zhang et al. 2024) regularizes this process with a preference objective, but still treats the complete QA pair as one optimization target. Together with their variants (Wang, Wei et al. 2025; Zhao et al. 2024; Wang et al. 2025; Yang et al. 2025), such objectives distribute forgetting pressure across many sequence positions, including syntactic function words, without explicitly separating knowledge-bearing content from linguistic structure. Consequently, stronger forgetting can coincide with degraded generation fluency and reduced general capabilities.

Token-level Unlearning Token-level methods narrow the target by suppressing specific token logits or redirecting them toward alternatives (Yu et al. 2025; Li et al. 2025a). Although this avoids some sequence-wide penalties, residual latent semantics may remain, while redirected generation can introduce factual hallucinations. Recently, Tan et al. (2025) identifies important tokens and suppresses attention toward them across the sequence. However, this strategy remains centered on static token salience rather than a context-specific query-to-key retrieval route, limiting contextual discrimination (Tran, Liu, and Xiong 2025; Jin et al. 2025). The same

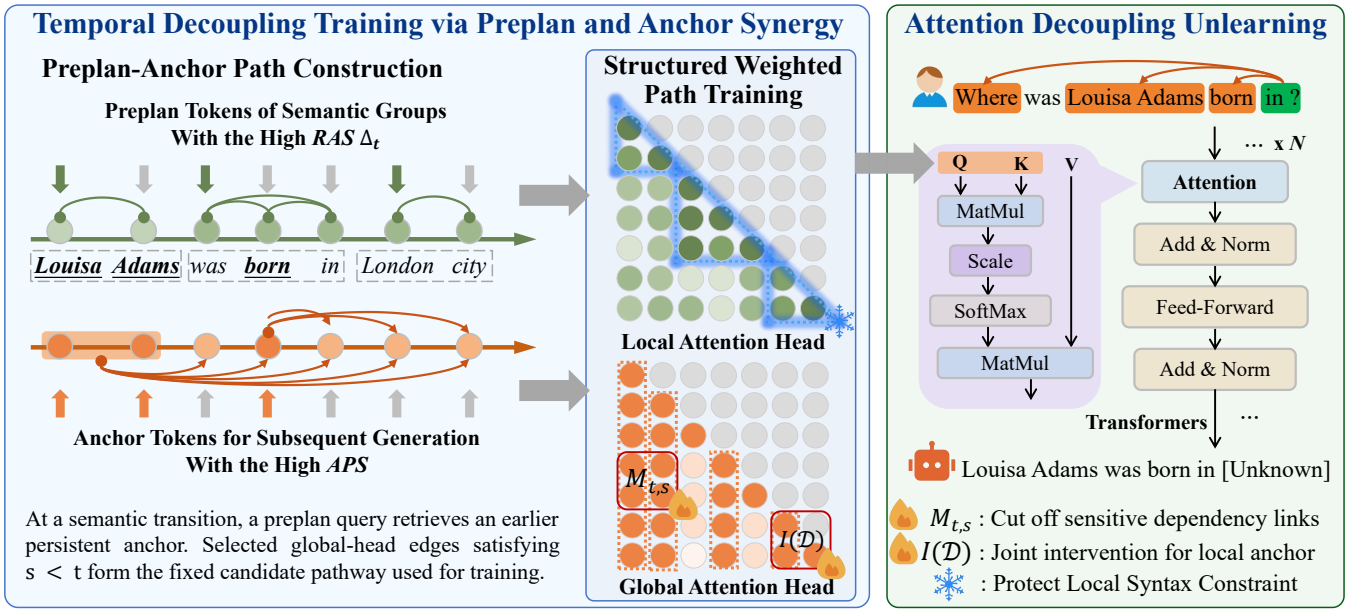


Figure 3: Illustration of the workflow. Before predicting the next important token, the preplan token "in" queries earlier sensitive anchors through long-range backward attention. After training, ADU suppresses this sensitive dependency path while preserving local anchor structure and maintaining the continuity of local semantic segments.

entity may therefore be weakened in both sensitive and benign contexts, causing excessive forgetting and disrupting normal knowledge retrieval (Yuan et al. 2025).

These families differ in granularity but define forgetting mainly over the sequence or token being generated, rather than the contextual computation that retrieves it. ADU instead identifies a recurring preplan–anchor rhythm under the original model and fixes the resulting candidate paths before training. Attention-projection adapters then suppress mass along these paths, while retain language modeling and local-attention preservation constrain collateral changes. Thus, ADU targets context-specific retrieval without directly erasing the sensitive token itself.

Method

ADU is a training-based unlearning method (see Figure 3). Given an original model θ_0, a forget set D_f, and a retain set D_r, it learns $\theta_1 = \mathcal{U}_{\mathrm{ADU}}(\theta_0; D_f, D_r)$ through attention-projection adapters. The learned pathway decoupling acts in every standard forward pass of θ_1, while counterfactual activation exchange is used only after training to identify the internal computation mediating forgetting.

Generation Inequality and Temporal Rhythm

During autoregressive generation, we formalize *generation inequality* as unequal temporal roles: local heads maintain short-range dependencies, whereas global heads retrieve distant, persistently attended tokens. Around semantic transitions, retrospective local shifts mark preplan positions, and persistent global attention identifies earlier anchors that shape generation. This pattern yields the preplan–anchor retrieval hypothesis operationalized below (Figure 2).

Let $x = (u, o)$ be a context continuation sequence and R_x the continuation positions. Let $A^{(l,h)}_{\theta,t,s}(x)$ be the causal attention weight from query position t to key position $s \le t$ in layer l, head h. On a held-out calibration split $\mathcal{D}_{\mathrm{cal}}$, the head's average backward distance under the model:

$$d^{(l,h)} = \mathbb{E}_{x\sim\mathcal{D}_{\mathrm{cal}}}\left[\frac{1}{|R_x|}\sum_{t\in R_x}\sum_{s\le t} A^{(l,h)}_{\theta_0,t,s}(x)(t-s)\right]. \quad (1)$$

The bottom and top ρ fractions form H_{loc} and H_{glob}, respectively. We use $\rho = 0.3$, compute the partition once from the original model, and keep it fixed during training.

Let $\bar{A}^{\mathrm{loc}}_{\theta_0}(x)$ and $\bar{A}^{\mathrm{glob}}_{\theta_0}(x)$ denote original-model attention averaged over the fixed local and global head sets. Given a retrospective distance cap W and a fixed future continuation window $F_x(s)$, we define

$$r_t = \sum_{s\le t} \bar{A}^{\mathrm{loc}}_{\theta_0,t,s}(x)\min(t-s, W),$$

$$\Delta_t = |r_t - r_{t-1}|, \qquad a_s = \frac{1}{|F_x(s)|}\sum_{u\in F_x(s)} \bar{A}^{\mathrm{glob}}_{\theta_0,u,s}(x). \quad (2)$$

Here, Δ_t is evaluated at continuation positions with a preceding continuation position, and positions without a valid future window are excluded from APS selection. Large Δ_t identifies a candidate transition, whereas large a_s indicates persistent influence on subsequent queries. For selection ratio q, let $Q_{1-q}(\Delta; x)$ and $Q_{1-q}(a; x)$ denote the corresponding within-sequence quantiles. We define

$$T_{\mathrm{pre}}(x) = \{t \in R_x \mid \Delta_t \ge Q_{1-q}(\Delta; x),\ r_t \ge \tau_{\mathrm{ras}}\},$$
$$S_{\mathrm{anc}}(x) = \{s \in R_x \mid a_s \ge Q_{1-q}(a; x)\} \cap C_f(x). \quad (3)$$

Here, $C_f(x)$ contains candidate sensitive positions derived from the forget continuation. Exact window construction and sensitive-position filtering are detailed in Appendix B. These signals locate candidate retrieval sites; ADU turns the identified rhythm into a trainable mechanism by suppressing the corresponding attention pathway.

Pathway Contributions and Causal Mediation

For each forget sample, $\mathcal{P}(x)$ contains tuples $e = (l, h, t, s)$ such that $(l, h) \in H_{\text{glob}}$, $t \in T_{\text{pre}}(x)$, $s \in S_{\text{anc}}(x)$, and $s < t$, where the last condition enforces causal attention order. We distinguish the contribution before and after the effective output projection:

$$\begin{aligned} \widetilde{C}_e(\theta, x) &= A^{(l,h)}_{\theta,t,s}(x) V^{(l,h)}_{\theta,s}(x), \\ C_e(\theta, x) &= W^{(l,h)}_{O,\theta} \widetilde{C}_e(\theta, x). \end{aligned} \tag{4}$$

Here, $\widetilde{C}_e$ is the head-space contribution manipulated by the intervention, while C_e is its residual-stream image controlled by the pathway objective. For $a \in \{0, 1\}$, define the path-specific causal mediator as

$$\mathbf{M}_a(x) = \left(\widetilde{C}_e(\theta_a, x) \right)_{e \in \mathcal{P}(x)}. \tag{5}$$

Replacing $\mathbf{M}_a(x)$ by $\mathbf{M}_b(x)$ subtracts the running model's selected contributions and adds the source model's corresponding contributions at each affected pre-output-projection head output. The recipient model retains its own W_O and every unpatched computation, and all downstream activations are recomputed. Here, $a = 0$ denotes the original model and $a = 1$ the trained ADU model.

For theoretical and causal analysis, let $I_f(x)$ denote the positions of a sensitive continuation span. For each $i \in I_f(x)$, let y_i be the target token and $\mathcal{C}_i(x)$ a prespecified set of non-target contrasts. Under teacher forcing, the sensitive accessibility score is

$$Y_\theta(x) = \frac{1}{|I_f(x)|} \sum_{i \in I_f(x)} s_\theta(y_i, x_{<i}), \tag{6}$$

where

$$s_\theta(y_i, x_{<i}) = z_\theta(y_i \mid x_{<i}) - \log \sum_{c \in \mathcal{C}_i(x)} \exp z_\theta(c \mid x_{<i}).$$

This dataset-agnostic score is used only to formalize sensitive accessibility and counterfactual effects; it is not part of the ADU training objective.

$$\begin{aligned} \Delta_f &= \mathbb{E}_{x \sim D_f} \left[Y_{\theta_0}(x) - Y_{\theta_1}(x) \right] \geq \epsilon_f > 0, \\ \Delta_r &= \mathbb{E}_{x \sim D_r} \left[\mathcal{D}_{\text{ret}} \left(\pi_{\theta_1}(\cdot \mid x), \pi_{\theta_0}(\cdot \mid x) \right) \right] \leq \epsilon_r, \end{aligned} \tag{7}$$

where $\mathcal{D}_{\text{ret}}$ measures retain-behavior discrepancy.

Let $Y(a, \mathbf{m}; x)$ denote the counterfactual score from θ_a when its selected head-space contributions are replaced by $\mathbf{m}$ before W_{O,θ_a}. All unpatched computations and parameters stay as θ_a, and downstream activations are recomputed. Define $Y_{ab}(x) = Y(a, \mathbf{M}_b(x); x)$, where the first index marks the running model and the second the mediator source. Writing $\mathbb{E}_f$ for expectation over D_f, we obtain

$$\begin{aligned} \text{TE} &= \mathbb{E}_f[Y_{00} - Y_{11}], \\ \text{IE}_{\text{sup}} = \mathbb{E}_f[Y_{00} - Y_{01}], &\quad \text{IE}_{\text{res}} = \mathbb{E}_f[Y_{10} - Y_{11}]. \end{aligned} \tag{8}$$

Under consistency, $Y(a, \mathbf{M}_a(x); x) = Y_{\theta_a}(x)$, so $Y_{aa} = Y_{\theta_a}$ and $\text{TE} = \Delta_f$. The suppression effect replaces Base contributions with their ADU counterparts, whereas the restoration effect replaces ADU contributions with their Base counterparts. These bidirectional interventions test whether the selected preplan–anchor contributions causally mediate the learned forgetting effect. Direct remainders representing additional unpatched routes are defined in Appendix A.

Training Objective and Theoretical Guarantees

Let $N_x = \max(1, |\mathcal{P}(x)|)$ and $A_{\theta,e}(x) = A^{(l,h)}_{\theta,t,s}(x)$ for $e = (l, h, t, s)$. The pathway mass and forget loss are

$$\text{PM}_\theta(x) = \frac{1}{N_x} \sum_{e \in \mathcal{P}(x)} A_{\theta,e}(x), \qquad \mathcal{L}_f = \mathbb{E}_{x \sim D_f}[\text{PM}_\theta(x)]. \tag{9}$$

The pathway indices are computed once under θ_0 and kept fixed during training; gradients flow through the selected attention values but not through the discrete mask. We combine $\mathcal{L}_f$ with language modeling on D_r and a row-wise loss that preserves original-model local attention on $D_f \cup D_r$:

$$\mathcal{L}_{\text{ADU}} = \alpha \mathcal{L}_f + (1 - \alpha) \left(\mathcal{L}_{\text{lm}} + \mathcal{L}_{\text{loc}} \right). \tag{10}$$

Here, $\alpha \in [0, 1]$ controls the forget–retain balance. The backbone remains frozen, while LoRA adapters update W_Q, W_K, W_V, and W_O in layers containing selected global heads. Samples with $\mathcal{P}(x) = \varnothing$ contribute zero to $\mathcal{L}_f$. The complete retain objective, trainable scope, and empty-path handling are detailed in Appendix B.

From pathway training to knowledge suppression. If $\|W^{(l,h)}_{O,\theta} V^{(l,h)}_{\theta,s}(x)\|_2 \leq B$ for all selected edges, then

$$\frac{1}{N_x} \sum_{e \in \mathcal{P}(x)} \|C_e(\theta, x)\|_2 \leq B \, \text{PM}_\theta(x). \tag{11}$$

Thus, minimizing pathway mass controls the average transported magnitude of selected edge contributions rather than treating attention weights as explanations. These contributions form the path-specific component of the attention-output computation examined by activation exchange.

For an affected query–head row j, let p_j and p'_j be its selected-anchor mass before and after pathway decoupling, with $\delta_j = p_j - p'_j \geq 0$. Write

$$\mathbf{o}_j(p_j) = p_j \mu_{S,j} + (1 - p_j) \mu_{\bar{S},j}, \qquad \Gamma_j = \mu_{S,j} - \mu_{\bar{S},j},$$

where $\mu_{S,j}$ and $\mu_{\bar{S},j}$ are the normalized selected-anchor and complementary value-output mixtures. A mass-transfer intervention holds these mixtures fixed while reducing p_j, yielding $\mathbf{o}_j(p'_j) - \mathbf{o}_j(p_j) = -\delta_j \Gamma_j$.

Let $g(\mathbf{p}) = Y(\mathbf{o}_1(p_1), \ldots, \mathbf{o}_J(p_J))$ be the sensitive score induced by the affected attention outputs. Assume that g is

Method	Forget tasks(%)		Retain tasks(%)		
Llama3.1-8B-Instruct	**Bio.↓**	**Cyber↓**	**MMLU↑**	**GSM8K↑**	**Flu.↑**
Base	71.86	45.37	68.16	67.83	3.74
NPO_KL[‡] (Zhang et al. 2024)	56.38	34.32	52.37	53.55	2.79
RMU[‡] (Li et al. 2025b)	39.55	31.75	53.43	52.64	2.90
ICUL[◇] (Pawelczyk, Neel, and Lakkaraju 2024)	41.31	30.46	60.69	58.40	**3.58**
ALU[◇] (Sanyal and Mandal 2025)	31.87	<u>29.94</u>	**63.78**	<u>58.65</u>	3.26
MET[†] (Yu et al. 2025)	34.63	31.47	52.19	54.19	3.18
ASU[†] (Tan et al. 2025)	34.49	33.17	58.58	57.24	3.31
ALTER[†] (Chen et al. 2026)	<u>29.69</u>	30.77	60.10	57.60	3.20
ADU[†] (Ours)	**27.32**	**27.97**	<u>62.84</u>	**58.82**	<u>3.34</u>
Qwen3-14B	**Bio.↓**	**Cyber↓**	**MMLU↑**	**GSM8K↑**	**Flu.↑**
Base	76.07	50.98	75.18	79.25	3.80
NPO_KL[‡] (Zhang et al. 2024)	60.59	39.85	64.88	63.50	2.88
RMU[‡] (Li et al. 2025b)	43.85	36.24	66.51	64.86	3.08
ICUL[◇] (Pawelczyk, Neel, and Lakkaraju 2024)	46.82	31.50	65.22	68.37	**3.69**
ALU[◇] (Sanyal and Mandal 2025)	<u>31.71</u>	<u>30.38</u>	67.67	68.81	<u>3.58</u>
MET[†] (Yu et al. 2025)	38.28	34.53	65.49	66.29	3.27
ASU[†] (Tan et al. 2025)	32.09	33.89	69.88	<u>71.54</u>	3.47
ALTER[†] (Chen et al. 2026)	34.24	36.58	**71.55**	70.83	3.35
ADU[†] (Ours)	**29.40**	**29.12**	<u>70.91</u>	**73.28**	3.57

Table 1: Multiple-choice accuracy on the forgetting/retention benchmark after unlearning. [†], [◇], and [‡] denote token-level training, prompt-based methods, and sequence-level training, respectively.

differentiable and, at every point along the intervention path, $\langle \nabla_{\mathbf{o}_j} g, \Gamma_j \rangle \geq \kappa_j > 0$. For the retain-discrepancy functional g_r, assume L-Lipschitz continuity and $\|\Gamma_j\|_2 \leq B_j$. Then

$$\begin{aligned} g(\mathbf{p}') - g(\mathbf{p}) &\leq -\sum_j \kappa_j \delta_j, \\ |g_r(\mathbf{p}') - g_r(\mathbf{p})| &\leq L \sum_j B_j \delta_j. \end{aligned} \tag{12}$$

The first inequality gives a sufficient condition for reducing sensitive log-odds, while the second bounds the retain-discrepancy change attributable to the same intervention. Complete proofs are provided in Appendix A.

Together, the pathway loss controls attention contributions, directional alignment translates their reduction into lower sensitive log-odds, and the retain objective constrains changes outside the pathway. Bidirectional activation exchange then tests whether the modified attention-output computation mediates the forgetting effect.

Experiment

Experiment Settings

Datasets We evaluate our method on three benchmarks. WMDP (Li et al. 2025b) verifies forgetting and retention effectiveness by assessing model knowledge in sensitive domains such as biosafety and cybersecurity. MUSE-Harry Potter (Shi et al. 2024b) assesses copyright unlearning: models are first fine-tuned on the Harry Potter book content to memorize it, then unlearned to forget that content. TOFU (Maini et al. 2024) examines boundary preservation between the forget set and its neighboring retain set, simulating synthetic and real-world unlearning scenarios. We note that all three benchmarks involve extended, context-rich responses where the model must retrieve factual knowledge through multi-token generation–precisely the regime where preplan-to-anchor attention patterns emerge and ADU's pathway decoupling is most effective. To test general ability, we apply MMLU (Hendrycks et al. 2021) for fact answering and GSM8K (Cobbe et al. 2021) for math reasoning. Datasets details and configurations are provided in Appendix C.1.

Metrics For WMDP, we report multiple choice accuracy on Bio and Cyber as forgetting metrics, where lower values indicate stronger forgetting. MMLU and GSM8K are used as retention metrics, where higher values indicate stronger utility preservation. For MUSE Harry Potter, we report BLEU and ROUGE-L to measure textual overlap with copyrighted content, together with MMLU and fluency for utility. For TOFU, we report ROUGE-L on the target unlearned data, Top 5 exclusion rate, Forget Quality, ROUGE-L on neighboring knowledge, neighboring accuracy, and general knowledge accuracy. Fluency is evaluated by GPT-4o on a 1 to 5 scale. We also report Forgetting Performance (FP) and Retaining Performance (RP) in analysis, where FP is the average of WMDP Bio and Cyber, and RP is the average of MMLU and GSM8K. Details are provided in Appendix C.2.

Baselines We compare with sequence-level training, prompt-based methods, and token-level training methods. Sequence-level training includes NPO_KL (Zhang

Method	TUD			NEK		GEK
Llama3.1-8B	**R-L↓**	**TR↑**	**FQ↑**	**R-L↑**	**Acc↑**	**Acc↑**
NPO_KL‡ (Zhang et al. 2024)	0.31	0.78	0.72	0.58	62.2	64.2
RMU‡ (Li et al. 2025b)	0.19	0.91	0.88	0.62	65.1	68.3
ICUL◇ (Pawelczyk, Neel, and Lakkaraju 2024)	0.17	0.94	0.54	0.61	63.8	69.0
ALU◇ (Sanyal and Mandal 2025)	0.13	0.95	0.67	0.64	66.7	70.6
MET† (Yu et al. 2025)	0.18	0.88	0.92	0.64	63.3	70.2
ASU† (Tan et al. 2025)	0.16	0.93	0.87	0.66	69.6	71.0
ALTER† (Chen et al. 2026)	0.14	0.91	0.81	0.60	67.2	71.3
ADU† (Ours)	**0.11**	**0.96**	**0.93**	**0.69**	**70.8**	**72.3**

Table 2: Performance comparison on TOFU (10%) with Llama3.1-8B-Instruct.

Method	BLEU↓	R-L↓	MMLU↑	Flu.↑
Original	74.80	85.14	46.33	3.63
NPO‡	**1.55**	14.08	42.70	2.96
WHP‡	23.68	17.93	43.49	2.52
ALU◇	7.21	14.85	44.34	3.27
ICUL◇	27.50	25.89	44.02	**3.34**
ALTER†	6.96	10.40	43.84	2.32
Ours†	4.78	**9.49**	**45.64**	3.29

Table 3: Results on MUSE-Harry Potter with Llama2-7B.

Setting	Avg↓	MMLU↑	GSM8K↑
ADU	27.65	62.84	58.82
w/o pathway loss	47.79	63.22	58.44
w/o retain objective	25.68	56.32	52.10
random heads	34.79	59.26	55.28
w/o anchor filter	26.19	58.40	54.32
w/o RAS preplan	32.38	60.05	55.99

Table 4: Component ablation on Llama3.1-8B-Instruct. Avg denotes the average of WMDP Bio and Cyber.

et al. 2024), and Representation Misdirection for Unlearning (RMU) (Li et al. 2025b). Prompt-based methods include ICUL (Pawelczyk, Neel, and Lakkaraju 2024) and ALU (Sanyal and Mandal 2025). Token-level training includes Model Edit Token (MET) (Yu et al. 2025), Attention Shift Unlearning (ASU) (Tan et al. 2025), and Hydra Suppress Unlearning (ALTER) (Chen et al. 2026).

Main Result

Forgetting-Retention Effectiveness We report WMDP forgetting and general retention results (Table 1). On Llama3.1-8B-Instruct, ADU reduces Bio accuracy from 71.86 to 27.32 and Cyber accuracy from 45.37 to 27.97, retaining MMLU and GSM8K at 62.84 and 58.82. On Qwen3-14B, ADU achieves the lowest Bio and Cyber accuracy and the best GSM8K retention among unlearning methods; prompt-based ALU ranks second on Cyber forgetting without modifying model parameters. These results show that ADU improves trainable forgetting and retention trade off instead of optimizing one forgetting metric at the cost of utility. Table 3 evaluates copyright unlearning on MUSE Harry Potter. ADU achieves the lowest ROUGE-L among unlearning methods and strongest MMLU retention. Although NPO gives lower BLEU, its MMLU drops to 42.70, while ADU keeps MMLU at 45.64 and fluency at 3.29. This pattern supports the pathway view. ADU weakens the route to memorized content while avoiding broad degradation of general next token behavior. Settings and costs are in Appendix C.4.

Boundary Preservation To further verify the impact of unlearning methods on neighboring and retained knowledge, we conducted experiments on TOFU (10%), as shown in Table 2. Sequence-level training methods struggle to balance the trade-off between forgetting and model utility. Prompt-based methods provide stronger retention, but their forgetting and neighboring preservation remain unstable. Although token-level training methods improve this balance, existing variants may still disrupt semantic dependencies shared with neighboring facts. For example, ASU obtains 69.6% NEK accuracy, but its TUD R-L remains 0.16. By severing hazardous retrieval pathways while preserving adjacent semantic pathways, ADU achieves the best TUD R-L of 0.11, TR of 0.96, and the best NEK and GEK of 70.8% and 72.3%, showing stronger boundary preservation and general utility.

Discussions

Component ablation Table 4 isolates each ADU component on Llama3.1-8B-Instruct. Removing the pathway loss raises forgetting metrics sharply, indicating that suppressing sensitive pathway mass drives forgetting. Removing the retain loss keeps forgetting but reduces MMLU and GSM8K by 6.52 and 6.72 points, showing that retain language modeling is essential for utility. Random heads weaken both forgetting and retention, suggesting that the local-global partition is non-interchangeable. Removing the sensitive anchor filter harms MMLU and GSM8K by penalizing benign high-APS anchors. Removing the RAS based preplan selection weakens forgetting and retention, confirming that ADU benefits from intervening at the transition point before sensitive an-

Condition	WMDP Avg.↓	MMLU↑
Base	58.62	68.16
Base − Selected $\widetilde{C}_e$	**36.37**	67.58
Base − Matched random $\widetilde{C}_e$	57.09	67.52
ADU	27.65	62.84
ADU ← Base selected	47.23	63.37
ADU ← Base matched random	28.82	61.64

Table 5: Path-specific contribution interventions on Llama3.1-8B-Instruct. "←" replaces selected contributions in the running model with their source-model counterparts.

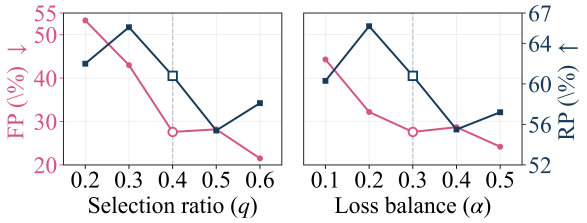


Figure 4: Parameter sensitivity analysis on WMDP and retention tasks with Llama3.1-8B-Instruct.

chors guide later generation. Full results including TOFU metrics are provided in Appendix D.2.

Path-specific causal validation. Figure 2(b) visualizes how RAS peaks mark preplan transitions and APS peaks identify persistently attended anchors, localizing candidate pathways without proving they control sensitive retrieval. We therefore intervene directly on the pre-output-projection contributions of the identified preplan–anchor edges. Removing them from Base lowers WMDP Avg. from 58.62 to 36.37, whereas removing a cardinality-matched random edge set yields 57.09, showing retrieval depends specifically on the selected pathway rather than an arbitrary same-sized perturbation. Conversely, replacing ADU's selected contributions with their Base counterparts restores WMDP Avg. from 27.65 to 47.23, while matched-random replacement reaches only 28.82. The selected interventions change MMLU by only -0.58 and +0.53 points, respectively. These complementary results link ADU's training target to its behavioral effect: the selected contributions support a substantial portion of Base retrieval, and restoring their original computation recovers much of the access suppressed by ADU. Appendix E provides complete bidirectional replacement analysis.

Hyperparameter Sensitivity Analysis We analyze the selection ratio q, and the loss balance α on Llama3.1-8B-Instruct. Figure 4 reports the forgetting-retention trade-off when varying one hyperparameter while fixing the others to their default values. Small q misses sensitive pathways and leaves higher WMDP accuracy, whereas large q includes benign anchors and harms retention. The default $q = 0.4$ achieves FP 27.65 and RP 60.83, which forms a stable trade-off knee while avoiding the retention degradation observed at larger q values. A small α underweights the pathway objective and generally weakens forgetting, whereas large α

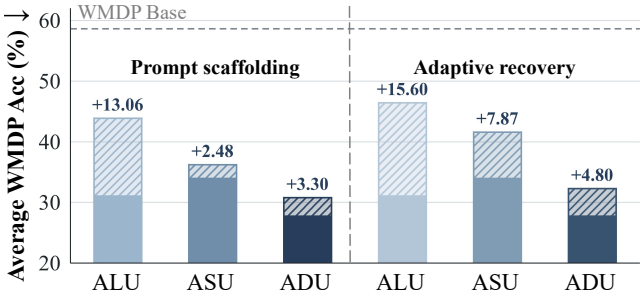


Figure 5: Groupwise worst-case WMDP accuracy under different attacks on Llama3.1-8B-Instruct.

provides limited forgetting gains and lowers retention performance. The default $\alpha = 0.3$ gives the best tested balance. Full numerical results and seed stability are reported in Appendix D.3 and Appendix D.4.

Robustness Analysis. We group six attacks into prompt scaffolding (few-shot, masking, and role-play CoT) and adaptive recovery (anchor shift, multi-turn probing, and repeated sampling). They test whether altered reasoning contexts or retrieval strategies re-elicit forgotten knowledge. As shown in Fig. 5, although ASU has a smaller increase under prompt scaffolding, ADU still achieves the lowest attacked accuracy. Under adaptive recovery, ADU has both the smallest increase and lowest final accuracy, remaining below ALU and ASU. Thus, pathway decoupling limits knowledge recovery beyond the original prompt. More results and details are in Appendix F.1.

Conclusion

Our work formulates LLM unlearning as contextual pathway decoupling: sensitive knowledge is retrieved through context-dependent internal routes and should not be reduced to an entire sequence, a static token, or a prompt-level refusal. Based on this view, we introduce ADU, which identifies a preplan–anchor rhythm from the temporal specialization of local and global attention heads, fixes candidate retrieval paths under the original model, and trains attention-projection adapters to suppress them while preserving retain-set language modeling and local-attention structure. The resulting framework provides a persistent parameter-level mechanism that targets sensitive retrieval in context, while limiting excessive forgetting, utility degradation, and recovery under altered prompts. Experiments on WMDP, TOFU, and MUSE-Books demonstrate strong forgetting–retention trade-offs, and bidirectional edge-contribution interventions establish that the selected paths causally mediate a substantial portion of sensitive retrieval and the learned forgetting effect. Future work will extend pathway identification beyond attention, improve automatic anchor construction, and develop stronger guarantees against residual knowledge recovery.

References

Bhaila, K.; Van, M.-H.; and Wu, X. 2025. Soft prompting for 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), 4046–4056.

Cha, S.; Cho, S.; Hwang, D.; Lee, H.; Moon, T.; and Lee, M. 2024. Learning to unlearn: Instance-wise unlearning for pre-trained classifiers. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 11186–11194.

Chen, X.; Guo, J.; Li, Y.; Wang, Z.; Gong, Y.; Zou, J.; Wei, J.; and Tian, W. 2026. ALTER: Asymmetric loRA for token-entropy-guided unlearning of LLMs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 35366–35374.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Geng, J.; Li, Q.; Woisetschlaeger, H.; Chen, Z.; Cai, F.; Wang, Y.; Nakov, P.; Jacobsen, H.-A.; and Karray, F. 2025. A comprehensive survey of machine unlearning techniques for large language models. *arXiv preprint arXiv:2503.01854*.

Gong, Q.; Yang, X.; Chen, X.; Lai, J.; Meng, H.; and Tang, X. 2026. FedOrtho: Efficient Federated Unlearning Via Orthogonal Convolution and Adaptive Soft Pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, 8009–8018.

Grynbaum, M. M.; et al. 2023. The Times sues OpenAI and Microsoft over AI use of copyrighted work. *The New York Times*, 27(1).

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*.

Hu, Z.; Zhang, Y.; Xiao, M.; Wang, W.; Feng, F.; and He, X. 2025. Exact and efficient unlearning for large language model-based recommendation. *IEEE Transactions on Knowledge and Data Engineering*.

Jiang, P.; Lyu, X.; Li, Y.; and Ma, J. 2025. Backdoor Token Unlearning: Exposing and Defending Backdoors in Pretrained Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 24285–24293.

Jin, M.; Luo, W.; Cheng, S.; Wang, X.; Hua, W.; Tang, R.; Wang, W. Y.; and Zhang, Y. 2025. Disentangling memory and reasoning ability in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1681–1701.

Kim, H.; Kim, K.; Chae, S.; and Yoon, S. 2026. Unlearning-aware minimization. *Advances in Neural Information Processing Systems*, 38: 93806–93829.

Lee, H. K.; Liu, R.; and Xiong, L. 2026. Direct Token Optimization: A Self-Contained Approach to Large Language Model Unlearning. In *Findings of the Association for Computational Linguistics: ACL 2026*, 42083–42100.

Lee, J.; et al. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.

Li, J.; Zhang, C.; Du, M.; Zhang, H.; Chen, Y.; Wei, Q.; Fang, J.; Wang, R.; Bi, S.; and Qi, G. 2025a. Forget the Token and Pixel: Rethinking Gradient Ascent for Concept Unlearning in Multimodal Generative Models. In *Findings of the Association for Computational Linguistics: ACL 2025*, 12179–12200.

Li, N.; Pan, A.; Gopal, A.; Yue, S.; Berrios, D.; Gatti, A.; Li, J. D.; Dombrowski, A.-K.; Goel, S.; Mukobi, G.; et al. 2025b. The WMDP Benchmark: Measuring and Reducing Malicious Use with Unlearning. In *International Conference on Machine Learning*, 28525–28550. PMLR.

Lin, Z.; Liang, T.; Xu, J.; Liu, Q.; Wang, X.; Luo, R.; Shi, C.; Li, S.; Yang, Y.; and Tu, Z. 2025. Critical Tokens Matter: Token-Level Contrastive Estimation Enhances LLM's Reasoning Capability. In *International Conference on Machine Learning*, 37906–37918. PMLR.

Liu, Z.; Maharjan, S.; Wu, F.; Parikh, R.; Bayar, B.; Sengamedu, S. H.; and Jiang, M. 2025. Disentangling biased knowledge from reasoning in large language models via machine unlearning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6105–6123.

Maini, P.; Feng, Z.; Schwarzschild, A.; Lipton, Z. C.; and Kolter, J. Z. 2024. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*.

Nguyen, T. T.; Huynh, T. T.; Ren, Z.; Nguyen, P. L.; Liew, A. W.-C.; Yin, H.; and Nguyen, Q. V. H. 2025. A survey of machine unlearning. *ACM Transactions on Intelligent Systems and Technology*, 16(5): 1–46.

Pawelczyk, M.; Neel, S.; and Lakkaraju, H. 2024. In-Context Unlearning: Language Models as Few-Shot Unlearners. In *International Conference on Machine Learning*, 40034–40050. PMLR.

Pu, J.; Shi, M.; Ren, X.; Wang, Y.; Zhang, X.; Wang, Z.; and She, K. 2026. Decoding-Unlearning: Fact Forgetting via Entropy-Guided Inference. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 39834–39860.

Ranjan, R.; Grover, U.; Lin, X.; and Polyzou, A. 2026. Razor: Ratio-aware layer editing for targeted unlearning in vision transformers and diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7998–8008.

Sanyal, D.; and Mandal, M. 2025. Agents are all you need for LLM unlearning. *arXiv preprint arXiv:2502.00406*.

Shah, R. S.; Huang, J.; Murugesan, K.; Baracaldo, N.; and Yang, D. 2025. The unlearning mirage: A dynamic framework for evaluating LLM unlearning. In *Second Conference on Language Modeling*.

Shi, D.; et al. 2024a. Large Language Model Safety: A Holistic Survey. *CoRR*, abs/2412.17686.

Shi, W.; Lee, J.; Huang, Y.; Malladi, S.; Zhao, J.; Holtzman, A.; Liu, D.; Zettlemoyer, L.; Smith, N.; and Zhang, C. 2025. Muse: Machine unlearning six-way evaluation for language models. In *International Conference on Learning Representations*, volume 2025, 27797–27818.

Shi, W.; Malladi, S.; Zhao, J.; Holtzman, A.; Liu, D.; Zettlemoyer, L.; Smith, N. A.; and Zhang, C. 2024b. MUSE: Machine Unlearning Six-Way Evaluation for Language Models. *arXiv preprint arXiv:2407.06460*.

Sun, H.; Zhu, T.; Chang, W.; and Zhou, W. 2025. Generative adversarial networks unlearning. *IEEE Transactions on Dependable and Secure Computing*.

Tan, C.; Qu, Y.; Li, X.; Zhang, H.; Cui, S.; Chen, C.; and Gao, L. 2025. Wisdom is Knowing What not to Say: Hallucination-Free LLMs Unlearning via Attention Shifting. *NeurIPS 2025*.

Thudi, A.; Deza, G.; Chandrasekaran, V.; and Papernot, N. 2022. Unrolling sgd: Understanding factors influencing machine unlearning. In *EuroS&P 2022*, 303–319. IEEE.

Tran, T.; Liu, R.; and Xiong, L. 2025. Tokens for Learning, Tokens for Unlearning: Mitigating Membership Inference Attacks in Large Language Models via Dual-Purpose Training. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 22872–22888. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.

Wang, L.; Zeng, X.; Guo, J.; Wong, K.-F.; and Gottlob, G. 2025. Selective forgetting: Advancing machine unlearning techniques and evaluation in language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 843–851.

Wang, Y.; Wei, J.; et al. 2025. LLM Unlearning via Loss Adjustment with Only Forget Data. In *The Thirteenth International Conference on Learning Representations*.

Wang, Z.; Guo, J.; Pu, J.; Pu, H.; Yang, M.; Chen, X.; Ou, J.; Li, W.; Luo, G.; and Tian, W. 2026. CAP: Controllable Alignment Prompting for Unlearning in LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Yang, T.; Dai, L.; Wang, X.; Cheng, M.; Tian, Y.; and Zhang, X. 2025. Cliperase: Efficient unlearning of visual-textual associations in clip. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 30438–30452.

Yu, M.; et al. 2025. UniErase: Unlearning Token as a Universal Erasure Primitive for Language Models. *arXiv preprint arXiv:2505.15674*.

Yuan, H.; Jin, Z.; Cao, P.; Chen, Y.; Liu, K.; and Zhao, J. 2025. Towards robust knowledge unlearning: An adversarial framework for assessing and improving unlearning robustness in large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 25769–25777.

Zhang, R.; Lin, L.; Bai, Y.; and Mei, S. 2024. Negative Preference Optimization: From Catastrophic Collapse to Effective Unlearning. In *First Conference on Language Modeling*.

Zhao, H.; Yuan, C.; Huang, F.; Hu, X.; Zhang, Y.; Yang, A.; Yu, B.; Liu, D.; Zhou, J.; Lin, J.; et al. 2025a. Qwen3guard technical report. *arXiv preprint arXiv:2510.14276*.

Zhao, K.; Kurmanji, M.; Bărbulescu, G.-O.; Triantafillou, E.; and Triantafillou, P. 2024. What makes unlearning hard and what to do about it. *Advances in Neural Information Processing Systems*, 37: 12293–12333.

Zhao, S.; Wu, X.; Nguyen, C.-D. T.; Jia, Y.; Jia, M.; Yichao, F.; and Tuan, L. A. 2025b. Unlearning backdoor attacks for llms with weak-to-strong knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2025*, 4937–4952.

Zhuang, H.; Zhang, Y.; Guo, K.; Jia, J.; Liu, G.; Liu, S.; and Zhang, X. 2025. SEUF: Is Unlearning One Expert Enough for Mixture-of-Experts LLMs? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8664–8678.